

Not quite shipshape

Japan is building a superb vessel for ocean drilling but seems reluctant to provide the necessary resources for using it. It is missing an opportunity to take a scientific lead.

The biggest and best scientific drilling ship ever is being built in Japan. It won't be ready for research until late 2006 at the earliest, but researchers around the world are already planning how to use it. The country had to dig deep to find the ¥60 billion (US\$550 million) for the ship, *Chikyu*, and the annual maintenance and operation costs are estimated at a further ¥6 billion.

The prospect of drilling holes down to the Earth's mantle is exciting scientists, but Japan is neglecting the more mundane preparatory seismic surveys of the sea floor. These surveys are essential for the scientific and operations committees of the Integrated Ocean Drilling Program (IODP), which allocate time on research drilling cruises, including those on the *Chikyu*. The survey data are used to judge the scientific merit, environmental impact and safety of proposals. Last week the IODP's science committee met in Japan to prioritize projects for 2005–6 (see page 795). These people are no pushovers; scrimping on survey data will not go unnoticed.

Even more troublesome is the neglect of human resources. When Japan became a full partner in the IODP in October 2003, the number of slots for Japanese researchers on each cruise rose from about 12 to 100 per year. But there have been no new faculty positions or student fellowships to exploit this opportunity. Japan is already having difficulty filling its slots on the committees that form the core of the IODP's decision-making. And Japan's only academic oceanographic research centre, the University of Tokyo's Ocean Research Institute, recently implemented a budget cut that cost five faculty positions.

Japanese researchers did well at last week's meeting. Two IODP projects to study the Nankai Trough were ranked second and third in

priority. These projects had been developed over several years with seismic studies included. They were finely honed. But researchers fear that similar groundwork is not being laid for future missions.

In industry, roughly 15% of the total drilling costs goes on site surveying. On that basis, Japan should spend at least US\$10 million per year. The United States, an equal partner in the IODP, spends at least this much, and some estimates put its annual support for various IODP activities at more than \$30 million per year.

Yet Japan currently has no funds allocated for seismic surveys, although the Japan Marine Science and Technology Center has funding for its own operations. Some researchers have proposed the creation of a funding framework like that in United States, but in the meantime, those hopeful of getting money for seismic surveys have to apply for competitive grants in an open field. The \$5 million or so needed for each survey is beyond the scope of such grants.

Japan has a history of buying expensive equipment, such as atomic force microscopes, but not of budgeting for the technicians needed to use it effectively. Frustrated researchers will tell you that the heads of their institutes and bureaucrats simply want to spend big budgets on devices they can show off, without following through to make sure that the science gets done. Too often, machines just sit idle.

This won't happen with *Chikyu*. Researchers from the United States, the European Union and China, among others, are competing for time on the ship. As last week's meeting showed, the slots will go to those who do their homework before coming to class. One way or another, Japan needs fully to support the science, so its scientists can take the leadership role they deserve. ■

PhD — club or history?

The decision by the University of Konstanz to retract the PhD of Jan Hendrik Schön is misguided.

There are good reasons why scientific misconduct tends not to be punishable in law. The distinctions between bad science, sloppiness, data manipulation and purposeful fraud are subtle and often difficult to establish. Fabricated results tend to be discovered, thanks to the self-regulating mechanism of research replication. Scientists will find their own means of punishment.

The case of Jan Hendrik Schön shows how effective this mechanism can be. The rise of the young German physicist at Bell Laboratories in New Jersey ended abruptly two years ago when the inability of competing groups to reproduce his results led to him being found to have fabricated data in at least 16 high-profile publications. Almost overnight, the former *wunderkind* became a *persona non grata*.

The University of Konstanz has now decided that this is not the end of the story. In a bid to repair the damage done to the reputation of science, it has withdrawn Schön's PhD, which he obtained there in 1997 (see *Nature* 429, 692; 2004). An internal investigation last year concluded that his thesis was free of false data. But the university says that Schön has abused his entrance ticket to the academic world in such a malicious and irresponsible way that he forfeited his academic consecration. The "dignity" of the doctorate is at stake, it says.

Some will applaud this move. It is a clear warning to all scientists that honesty and good practice are life-long duties. The University of Konstanz, after all, has a right to dissociate itself from fraudsters taking advantage of its name. And when it comes to scientific misconduct, surely we need a policy of zero tolerance.

Well, yes, up to a point. But there is a sense that scientists are taking revenge for a particularly high-profile piece of misconduct that cast a wide shadow. Is Schön being made a scapegoat for the community's broader imperfections? And is it more reprehensible to cheat in physics than in biomedicine? After all, no one ever questioned the PhDs of Germany's most notorious science fraudsters, cancer researchers Friedhelm Hermann and Marion Brach, who fabricated data in about 100 papers.

At the heart of the issue is a question about the status of a PhD. The dignity of the degree is more threatened by the tens of thousands of fake PhDs available for sale than by one researcher going astray. No one doubts that Schön obtained his doctorate for a scientific achievement. The decision to revoke it implies that the PhD is a club membership that can be withdrawn. *Nature*, on the other hand, sees the PhD as a piece of history that cannot, and should not, be rewritten. ■





Tourist trail

Launch fails to dispel doubts over private space flight
p792



Planet plans

Experts call for NASA reform to secure martian success
p793



SUPA power

Scottish physicists gear up for collaboration
p794



Physics leader

Nobel laureate lands top job at Berkeley laboratory
p796

Biologists seek stamp of approval to send live fruitflies by post

David Cyranoski

Biomedical researchers are not normally viewed as outlaws. But those who post live creatures, such as fruitflies (*Drosophila*), to their colleagues are currently breaking an international agreement.

So *Drosophila* distributors have launched a bid to get international postal regulations modified. If they succeed, it would be unambiguously legal to send live fruitflies through the post.

The bid is backed by the US Department of State and will be considered at this September's meeting of the Universal Postal Union (UPU) — the United Nations organization that administers global postal rules — in Bucharest, Romania. Officials close to the process say they are confident that the amendment will be approved.

The contents of mailbags are coming under ever-closer scrutiny, and the need for fruitflies in the lab is on the rise. But many researchers and materials distributors around the world are a little hazy about the existing rules on sending live organisms through the post. The current Universal Postal Convention states that only bees, leeches, silkworms and “parasites and destroyers of noxious insects intended for control of those insects” are exempt from its bar on the international posting of live animals. Research organisms are out of luck, says Akhilesh Mathur, mail programme manager of the UPU, which administers the convention.

So Kevin Cook, who manages the Bloomington *Drosophila* Stock Center at Indiana University, is leading a drive to change the rules. He points out that *Drosophila* use is now central to biological and medical research worldwide, with close to 5,000 *Drosophila*-related papers published every year. Thousands of researchers have benefited from his centre's services, he says, including those at 421 laboratories in 38 countries to which the centre sent 43,000 fly cultures in 2003 alone. Cook sees the amendment as necessary to ensure that this trade can continue within international rules.

With the assistance of the state department, Cook is rallying support for the convention to pass an amendment that would add



Mail orders: current international rules prohibit the transfer of most live animals by post.

“flies of the family Drosophilidae for biomedical research” to the current list of exemptions.

Although researchers have not complained much about the restriction — whose level of enforcement is difficult to establish — Cook says that it needs to be changed because authorities have been paying much closer attention to the precise contents of mail, especially in view of the bioterrorism scares that followed the terrorist attacks of 11 September 2001. The idea for the change arose in “the uncertain political context of mailing something that doesn't conform to postal regulations”, he explains. Strict enforcement of the Cartagena Protocol on Biosafety, which sets restrictions on the shipment of genetically modified organisms, also threatens to restrict the movement of *Drosophila* (see *Nature* 428, 6; 2004).

Delivery service

But some scientists and distributors seem to think that posting *Drosophila* and other organisms is already allowed. The homepage for Kyoto Institute of Technology's *Drosophila* Genetic Resource Center, Japan's largest distributor, for example, states that “perishable biological substances” are exempt from postal restrictions.

And officials at the *Caenorhabditis*

Genetics Center at the University of Minnesota, a major international distributor of the nematode *Caenorhabditis elegans*, another ubiquitous model organism for research, say that they have experienced few difficulties when labelling the worms as “non-pathogenic, non-parasitic biological specimens of no commercial value” for the roughly 35 international shipments they make a week. “When the authorities ask for additional information we tell them exactly what's inside, and include publications showing their use in research,” says Theresa Stiernagle, curator of the centre.

Mathur says that the proposal is likely to be passed by the convention, which only takes place every five years, in which case it would come into effect on 1 January 2006. But even if this happens, researchers say privately that the change could create headaches for shippers of other model organisms, by highlighting the fact that their shipment is against the rules.

If it fails, the implications could be severe for fruitfly researchers. Distributors might have to resort to the special shippers they already use for larger organisms, increasing their costs tenfold. In the worst case scenario, Cook says, “the stock centre would stop sending flies to anyone outside the United States because the costs would be too high.”

Experts challenge claims for space tourism

Helen Pearson, Mojave

The sun was fierce and expectations were sky high when SpaceShipOne took flight over the windswept Mojave desert early on Monday morning. The rocket has become the first privately funded, manned spaceship to skim the edge of space. But experts say its creators may have to go back to the drawing board if they want a craft capable of ferrying passengers into orbit.

Enthusiasts claim that an era of affordable, commercial space travel will be ushered in by the \$20-million spacecraft, which was designed by aerospace engineer Burt Rutan and his company Scaled Composites, and backed by Microsoft co-founder Paul Allen. The launch is also Rutan's stepping stone towards the Ansari X prize, a \$10-million

jackpot for the first privately financed rocket to send three passengers, or an equivalent weight, up to a height of 100 kilometres twice within a fortnight.

Although aerospace engineers have applauded Rutan's feat, they say there are still numerous technological and financial hurdles to be overcome. "It's a wonderful achievement — but whether it leads to long term space tourism is really questionable," says Jerry Grey, director for aerospace policy at the American Institute of Aeronautics and Astronautics in New York. "It is insignificant in the overall scheme of space flight," says space-history expert Roger Launius of the National Air and Space Museum in Washington DC.

On 21 June, SpaceShipOne piggy-backed

on a carrier aeroplane to about 15 kilometres. Once released, it fired its custom-built hybrid engine, powered by a solid mix of synthetic rubber and nitrous oxide, or laughing gas, to reach 100 kilometres. This fuel cocktail is less flammable than the Space Shuttle's mix of liquid oxygen and liquid hydrogen, making for a safer, though less punchy, blast.

The craft re-entered the atmosphere by flipping up the back part of its wings into a high-drag configuration that is designed to ensure it sinks slowly without generating too much heat. This is different from conventional rockets, which use a computerized system to keep them at the right angle during descent. Ninety minutes after take off, pilot Mike Melvill landed safely back on Earth.

Still, not everything went as planned. The rocket crested at about 100.1 kilometres, but Rutan hoped to reach nearly 110 kilometres. It strayed over 30 kilometres off course when it re-entered the atmosphere. And the ship also touched down with a dent in its tail.

To reach continuous orbit some 300–400 kilometres above Earth will take top speeds that are eight times as fast as that of SpaceShipOne and an exponential leap in efficiency and energy, experts say.

Rutan would not confirm whether he plans to build orbital craft. But some of the other 26 X-prize contenders, and a few companies outside the competition, are already testing prototype spaceships with orbital flight in mind. SpaceX in El Segundo, California, for example, is scheduled to put a \$30-million US Department of Defense satellite into orbit on a prototype vertical-take-off rocket later this year.



Down to Earth: SpaceShipOne reached the runway safely, but it failed to stick to its planned course.

RNA therapy beckons as firms prepare for clinical trials

Erika Check, Washington

Biotechnology companies are edging closer towards taking a promising gene-silencing technique called RNA interference (RNAi) into clinical trials.

The technique can be used to inhibit viruses in animals, such as hepatitis in mice. But it has not yet been tested in people, although companies are consolidating their intellectual property positions and filling their bank accounts in the hope of launching clinical trials, possibly this year.

RNAi relies on a natural mechanism in which animal cells target and destroy foreign genetic material (see *Nature* 425, 10–12; 2003). Injecting small pieces of RNA into a cell seems to trigger the destruction of any pieces of RNA matching the injected sequence. This creates the enticing prospect of 'knocking down' expression of RNA in viruses such as HIV or

hepatitis, as well as correcting neurological problems or muscle disease.

One firm pioneering the approach was founded by biologist and Nobel laureate Philip Sharp of the Massachusetts Institute of Technology. His company, Alnylam Pharmaceuticals, based in Massachusetts, went public in May, raising \$30 million.

Alnylam merged last year with Ribopharma, a company based in Kulmbach, Germany, which owns a European patent on RNAi. It is working in diseases such as macular degeneration (progressive vision loss) and aims to test an RNAi technology in AIDS patients with the condition next year.

Benitec, a firm based in Queensland, Australia, has emerged as Alnylam's main rival. Last year, the US Patent and Trademark Office granted Benitec a patent that seems to

cover a large range of clinical applications of RNAi. And the firm has just acquired the company Avocel, founded by Stanford geneticist Mark Kay. Co-founder Sara Cunningham says the company hopes to be listed on the US NASDAQ stock exchange by the end of the year.

Benitec also plans to fund a clinical trial for geneticist John Rossi at the City of Hope in Duarte, California. Earlier this month, Rossi said he hoped to begin as early as this July. His trial will aim to treat children infected with HIV. The children's bone-marrow stem cells will be removed and treated with an RNAi molecule. The cells will then be implanted back into the children. It is hoped that the treated cells will resist HIV infection better than untreated cells. This could be the first RNAi clinical trial to be submitted to the US Food and Drug Administration.



Mission commission: Bush's panel recommends turning NASA labs into federally funded centres.

NASA reforms needed to give Moon–Mars plan a better shot

Tony Reichhardt, Washington

President Bush's ambitious plan to send astronauts to the Moon and Mars will only get off the ground if NASA centres are run differently, says a commission appointed to look into the plan's execution.

On 16 June, the presidential commission published a report calling on NASA to give private contractors or universities the management of its network of ten major space centres, such as the Goddard Space Flight Center in Maryland.

But critics say that the idea faces political obstacles that could doom it from the start. Plans to involve the private sector more deeply in manned space flight amount to little more than wishful thinking, they charge — despite this week's independent foray into space (see opposite page).

Bush laid out his plan in January (see *Nature*, 427, 183; 2004) and asked the panel, led by former US Air Force secretary Edward 'Pete' Aldridge, to determine how it might be carried out. The commission's report supports Bush's goal of sending humans beyond Earth orbit, but says that "the infrastructure of NASA is too large".

The panel had considered how to shut down some of the centres and Senator Sam Brownback (Republican, Kansas), who chairs the Senate subcommittee that oversees NASA, expressed interest in a law that would pave the way for such closures.

But Aldridge says politicians from states with NASA centres would almost certainly have blocked such a move. "Our report would probably have been burned on the first day," if it had taken that approach, he told a Senate hearing on 17 June.

Instead, the commission recommended that NASA convert its centres to Federally Funded Research and Development Centers (FFRDCs), like its Jet Propulsion Laboratory in Pasadena, California. The scientists and engineers at such centres are employees of universities or industrial companies, rather than civil servants.

But Paul Light, who studies federal workforce issues at the Brookings Institution in Washington, doubts that this would work for most NASA centres. "There's no magic in conversion to FFRDC status," he says. Light adds that the concept is sometimes just a ruse to move civil servants off the federal payroll without actually reducing costs or employee numbers.

And Marc Cohen, a space project planner and head of the union of scientists and engineers at NASA's Ames Research Center in California, warns that NASA scientists could fare badly under such a plan. He says it would favour centres that can do contract engineering work. "Most of us work very far away from the marketplace," Cohen says.

The Aldridge commission also said that "an independent space industry does not really exist" in the United States yet, and that the private sector was unlikely to put money into a Moon–Mars programme. But it added that there were steps that NASA could take to promote independent investment, such as offering cash prizes to innovators.

Senator Brownback agrees: a NASA authorization bill he introduced last week would give an award to the first outside group to put a person into orbit, and would call for the agency's next robotic lunar mission to be privately operated. ■

British drug company to put data online as criticism mounts

Jim Giles, London

Things are hotting up for Britain's largest drugs firm, GlaxoSmithKline (GSK), in the wake of a legal challenge launched earlier this month by Eliot Spitzer, New York state's attorney-general.

On 18 June, the British parliament announced an investigation into the pharmaceutical industry. On the same day, London-based GSK said that it would make all clinical trial data for marketed drugs available online to the public. But two days later, British lawyers representing patient groups announced fresh legal action against GSK.

The company's online data registry, due to launch later this year, will include summaries of trial protocols and data for everything from initial toxicology and safety studies to trials run on drugs after they have been licensed. "All the important results will be there," says GSK spokesman David Mawdsley.

Results from one such post-licensing trial are important to Spitzer's case. The attorney-general alleges that GSK failed to disclose data showing that the antidepressant paroxetine, known as Paxil in the United States and Seroxat in Britain, is no more effective than a placebo for depressed young people and might even increase the risk of suicide in this group (see *Nature* 429, 589; 2004). The drug has been approved by US regulators for use in adults, but it is frequently prescribed to under-18s.

Mawdsley says they have been considering the registry for a while: "This is going to be a big job. We started discussing it well before the Spitzer case."

GSK is also expected to hear shortly from Hugh James, a Cardiff law firm which says it is acting on behalf of 3,500 patients who claim to have suffered withdrawal symptoms when coming off Seroxat. The company, which says it is also representing relatives of "a few tens" of people who committed suicide while on the same drug, say they will outline their claim in a letter to GSK in the next few weeks. GSK declined to comment.

Later this year, the drug manufacturer may have to defend itself on a third front, when the House of Commons opens an inquiry into the pharmaceutical industry. The wide-ranging investigation, due to begin in September, will look at the influence of the industry on medical research, the promotion of drugs and regulatory reviews of drug safety and efficacy. ■

Biologists take degrading route to tackle cancer

Alison Abbott, Munich

Not content with genomes and proteomes, biologists have added another ‘-ome’ to their lexicon: the degradome. Its proponents say that it signals a fresh approach to cancer research. And the European Commission seems to agree — on 15 June it awarded one of the largest-ever grants for life-sciences to a four-year project based on the concept.

The degradome encompasses all of the enzymes in the body that degrade proteins. There are about 560 such enzymes, called proteases, in humans. Many of them modify the environment of human tissues and contain their growth — making sure a liver stays liver-shaped, for example.

The range and levels of proteases expressed by a tissue change dramatically if it becomes cancerous. For example, normal tissue cells are fixed in position by a matrix. But cancer cells seek to break this down using certain proteases, so that they can escape into the blood stream and establish themselves in different tissues. Other proteases are marshalled to help build the blood vessels needed to feed growing tumours.

But therapeutic approaches based on disrupting such protease activity have so far been unsuccessful in clinical trials. Dylan Edwards, a cancer biochemist at the University of East Anglia in Norwich, UK, and coordinator of the €10.4-million (US\$12.6-million) project, says that this is because the drugs tried have been designed to inhibit as many proteases as possible.

“Dogma has insisted that protease action must be blocked in as broad a manner as possible in cancer,” says Edwards. “But we now know that the web of proteases is complex. Some of them inhibit, rather than facilitate, cancer growth.” So researchers in the European project plan to investigate the biology of proteases and then design specific inhibitors to block them individually.

Involving 41 scientists from 13 countries, the project will cut across many different disciplines. It will depend initially on two mouse models — for breast and prostate cancer — later expanding to include colorectal and skin cancers. Biologists will study the expression of proteases to untangle which protease does what in normal and diseased tissues. The chemists in the team will work on developing highly selective inhibitors for diagnostics and therapeutics. ■

► www.cancerdegradome.org



Physicists at Edinburgh's James Clerk Maxwell building will forge alliances with Scottish colleagues.

Scots propose SUPA plan for united approach to physics

Jim Giles, London

Scotland's top physicists are planning to band together to create a single national physics department. Advocates say that the plan will help them to compete with the world's elite research institutions.

Under the scheme, which is part of a wider discussion about pooling resources across Scottish science, the country's six highest-rated physics departments will function as a single unit — the Scottish Universities Physics Alliance (SUPA). Members would remain in their individual universities but would bid for big grants together and offer joint graduate courses. Undergraduate teaching would remain outside the collaboration.

“This will brand Scottish physics with a strong identity,” says Alan Miller, the vice-principal for research at the University of St Andrews, which is a partner in SUPA. “We want to compete with Oxford, Cambridge and even the Massachusetts Institute of Technology and Harvard.”

The idea, which will be considered by the Scottish Higher Education Funding Council on 25 June, would involve the creation of about 12 new positions at St Andrews and the other SUPA universities — Glasgow, Edinburgh, Heriot-Watt, Strathclyde and Paisley. Miller puts the cost at £30 million (US\$55 million) over five years, but says that it would pay for itself in the long-term by generating extra research income.

Some universities are likely to take leading roles in certain subjects, says Miller. For

example, Glasgow and St Andrews have strengths in particle physics and materials, respectively. But he rules out the possibility of closing down groups or moving researchers to a single physics building at one site.

The move reflects changes in England, where, for example, the University of Manchester and UMIST (the University of Manchester Institute of Science and Technology) are merging to create a stronger, single institution, due to open in October. Officials at University College London and Imperial College London flirted with a merger idea in 2002, before backing down in the face of staff protests (see *Nature* **420**, 350; 2002). In both cases, the need to become more globally competitive was the driver for change.

The trend may also be repeated in other subjects in Scotland. On the east coast, the chemistry departments at the universities of Edinburgh and St Andrews are considering merging their research and graduate teaching functions. A similar unit may be formed in the west of the country by the chemistry departments at Glasgow and Strathclyde.

“We’d welcome this,” says Sean McWhinnie, science-policy manager at the Royal Society of Chemistry in London. “In Manchester, the merger will produce a world-class chemistry department. The same would apply in Scotland.”

Council officials say that it is too early to judge whether the approach should be applied more widely, but add that they are discussing the possibility with regard to biomedical research and the Earth sciences. ■

Climate researcher takes academy hot seat

Geoff Brumfiel, Washington

Ralph Cicerone, an atmospheric chemist and leading climate-change researcher, is set to become the next president of the US National Academy of Sciences.

The academy's council announced the nomination of Cicerone, who is currently chancellor of the University of California, Irvine, on 15 June. Once elected — usually a formality under the academy's system — he will serve a six-year term from July 2005, replacing the current president, molecular biologist Bruce Alberts.

The 2,000-member academy is generally considered to be the United States' pre-eminent scientific body, and is regularly asked by the government to provide scientific advice on issues ranging from stem cells to global warming. The 61-year-old Cicerone will be the first climate-change scientist to become its president.

His nomination comes after a period in which the scientific community has clashed with the federal government over climate change. In 2001, Cicerone led an academy study that angered the Bush administration by confirming the likelihood that human activities are causing global warming.

Cicerone is certainly no stranger to politically sensitive research. In the early 1970s, he was among the first atmospheric chemists to tie chlorine to the destruction of the ozone layer, according to Sherwood Rowland, a chemist at UC Irvine who won the 1995 Nobel Prize in Chemistry for his work on ozone depletion. Cicerone testified at the



Top job: Ralph Cicerone will succeed Bruce Alberts in July 2005.

time about the impact of chlorofluorocarbons (CFCs) on the ozone layer.

Colleagues say that Cicerone's easy-going personality and skill as a consensus-builder have helped to maintain his credibility with policy-makers. "He's very practised in the relationship between scientific discovery and sound public policy," says James Anderson, an atmospheric chemist at Harvard University in Cambridge, Massachusetts, who worked with Cicerone at the University of Michigan.

The 2001 climate-change report was requested by President Bush just before a trip to Europe where he sought to repair damage done by the US withdrawal from the Kyoto Protocol on climate change (see *Nature* 411, 725; 2001). "There were people who thought

our panel would be seen as carrying the water for somebody's agenda," Cicerone says. But in the end, the group delivered a scientifically precise document that provided little support for the administration's position. "It was a very honest report," says Eileen Claussen, president of the Pew Center on Global Climate Change in Washington DC.

As chair of the 2001 study, Cicerone did "exactly what we want a president to be able to do," says Peter Raven, head of the Missouri Botanical Garden in St Louis and chair of the academy's nominating committee. Raven says that Cicerone's proven ability as a fundraiser and manager during his six years at UC Irvine also strengthened his candidacy.

As well as addressing politically awkward scientific issues, Cicerone pledges continued attention to education and international outreach — areas emphasized by Alberts during his two six-year terms. The academy produced a set of standards in 1996 that has strongly influenced science education in US schools, and Alberts helped form the InterAcademy Council, an international consortium of science academies, in 2000.

But it is Cicerone's experience in tackling tricky areas of science policy that most excites his fellow academy members. "He's quite experienced on Washington affairs," says Frank Press of the Washington Advisory Group, who served as president of the academy from 1981 to 1993. "He is a balanced, unbiased person, and that's what the academy needs these days." ■

Last-ditch funding keeps ocean drilling project afloat

David Cyranoski, Yokohama

The world's largest ocean-drilling research team was given a crucial injection of cash last week, saving the programme from losing its only dedicated drilling ship.

The scientific planning committee of the Integrated Ocean Drilling Program (IODP) met in Yokohama to rank 19 science proposals, each one vying for drilling time during the fiscal year starting in October 2006. But the committee faced a serious problem — not having enough money for a ship to host these projects.

The IODP's current ship, the *JOIDES Resolution* — which is provided by the United States as part of its contribution to the multinational project — was assured of funding only until May 2005.

The IODP would face heavy competition

and high prices if it then tried to commission a replacement ship, says Frank Rack, director of drilling programs for the US Joint Oceanographic Institutions, because high oil prices are driving oil companies to rent drilling ships for exploration.

The meeting was close to becoming nothing more than "an educational exercise," says Mike Coffin, a geophysicist at the University of Tokyo and chair of the planning committee.

But representatives of the National Science Foundation, which funds US involvement in the programme, announced they would provide enough money to keep the *JOIDES Resolution* afloat. The amount is not clear and will depend on the cost of the proposals selected, but it came as a relief to scientists who feared that the projects they

were ranking would be lost to a gap in funding. Instead, the committee now has to rush through plans for two additional cruises in 2005 and a possibility of six in 2006.

"We have to scramble," says Terrence Quinn, a committee member from the University of South Florida. "You don't want to have a boat without projects to go on it."

The top three projects picked are all geophysics studies — something of a surprise given the growing presence of climate studies in ocean drilling projects, says Coffin. The top priority is to drill through the Guatemala Basin, all the way through the ocean crust, revealing details of magma formation, hydrothermal effects and tectonic processes. Investigations into the earthquake-generating zone in the Nankai Trough off Japan are next on the list. ■

Fraud conviction casts doubt on study into prayer-power

San Francisco Exposure of an author's fraudulent history has prompted concerns about a paper that claims Christian prayer doubles the success rate of *in vitro* fertilization (IVF). The paper has now been removed from the website of the *Journal of Reproductive Medicine*.

Daniel Wirth, a Californian lawyer and psychic researcher, and Rogerio Lobo and Kwang Cha, both at Columbia University in New York at the time of the research, found that women undergoing IVF were twice as likely to become pregnant if they were prayed for by people in a different country (K. Y. Cha, D. P. Wirth and R. A. Lobo *J. Reprod. Med.* **46**, 781–787; 2001).

Earlier this year, Wirth pleaded guilty to defrauding the Pennsylvania cable company Adelphia Communications of more than \$1 million. Many scientists then began to question the validity of the data in his paper, especially as it also emerged that Wirth was alleged in those proceedings to have used false identities, and has authored many pieces of research claiming to prove paranormal activities. Wirth is now under house arrest in California.

Donna Kessel, managing editor of the *Journal of Reproductive Medicine*, says that the paper was peer-reviewed but that she is now carrying out an investigation into the study. Columbia University is also looking into the work.

Nobel laureate takes the reins at Berkeley laboratory

Washington Nobel-prizewinning physicist Steven Chu has been named as the next director of Lawrence Berkeley National Laboratory in California.

Chu, the first Asian-American to run a US Department of Energy laboratory, comes from Stanford University to replace retiring director Charles Shank.

The University of California, which manages the lab for the government, hopes



Incoming: Steven Chu is returning home to the Lawrence Berkeley National Laboratory.

Brazil becomes fresh haunt for ghost shark



San Diego Brazilian biologists say that they have found a new species of 'ghost shark' or chimaera.

Ichthyologist Jules Soto and his colleagues at the Vale do Itajai University in Santa Catarina first spotted the new fish in photographs taken by their students on a fishing boat off the coast of Rio de Janeiro in 2001. They spent three years finding more samples and characterizing the

species (above) and have submitted their report to the journal *Zootaxa*.

The fish are relatives of sharks and can grow up to 2 metres long. They live at depths of 500 metres and are thought to have changed little in 150 million years. More than 30 species of chimaera have so far been described. They have been found in most waters around the world, although none had been sighted off Brazil.

that Chu will help it win a contract to continue to run the lab, says the university's president, Robert Dynes. Following a spate of management and security scandals at federal labs, the US government invited fresh bids for contracts to manage Lawrence Berkeley and two nuclear-weapons labs — Los Alamos National Laboratory in New Mexico and Lawrence Livermore National Laboratory in California.

Chu won a Nobel prize in 1997 for using lasers to cool atoms while at Bell Laboratories in Murray Hill, New Jersey, and Stanford. But he began his career as a graduate student at the Berkeley lab. "To me, it feels like a homecoming," he says.

Flow chart plots water loss for Colorado river

San Diego The Grand Canyon area of the United States, through which the Colorado River flows, is experiencing one of the worst droughts in 500 years, hydrologists say.

The Colorado and its side rivers are the primary source of water for millions of people in the densely populated Los Angeles and San Diego areas of southern California. But water flow from the Colorado basin has been below average for five years in a row.

Using observations and tree-ring measurements, scientists at the US Geological Survey (USGS) have reconstructed the Colorado River's water flow for the past half millennium. They found that the average annual flow from 1999 to 2003 was almost twice as low as it was during the dry Dust Bowl years of the 1930s, and even lower than during the severe drought of 1590–94.

The drought is probably a symptom of climate change, says the report's lead author Robert Webb, a hydrologist for the USGS in Tucson, Arizona. "It could end very soon, but it could also last for 30 years — and then we would really be in trouble."

Collaboration powers up drive to archive digital data

Washington How can the burgeoning mass of digital records found everywhere from labs to schools be preserved? Those with answers should apply to the US National Science Foundation (NSF), which has set aside about \$2 million to address the issue.

NSF officials say that most people don't consider long-term storage when they copy data onto compact discs or hard drives. Some ambitious projects to create records of digital information are already under way, including the non-profit Internet Archive in San Francisco, which stores some 30 billion web pages on conventional computer servers. But such storage is vulnerable to accidents and erratic funding, officials say.

The grants, issued in conjunction with the Library of Congress and announced on 16 June, are available for researchers working on the hardware, software and management issues for preserving digital records.

♦ www.digitalpreservation.gov

Historian lands leading role for Muslim science

London The next head of the Organization of the Islamic Conference (OIC) — the main intergovernmental body of Muslim countries — has pledged to improve cooperation between researchers from Muslim countries and others worldwide.

Ekmeleddin Ihsanoglu was elected as secretary-general of the 57-member OIC at its annual meeting of foreign ministers in Istanbul last week. Ihsanoglu, who specializes in the history of science during the Ottoman Empire, is also president of the International Union for the History and Philosophy of Science.

He will hold the OIC position, based at its secretariat in Jeddah, Saudi Arabia, for four years, starting in January 2005.

One slip, and you're dead...

The lethal toxins produced by cone snails are in hot demand for neuroscience research, and are being developed as potent drugs. Laura Nelson visits a would-be snail 'farmer', for whom milking time is fraught with danger.

Jon-Paul Bingham fumbles around for a condom. Big Bertha is waiting. There's an awkward pause. "It has to be the non-lubricated kind," he says. Bingham rips open the packet and slips the prophylactic over a small plastic test tube.

Big Bertha is one of Bingham's nine tropical marine cone snails. These colourful creatures are some of the most venomous beasts on the planet. But the powerful poisons they produce can, in tiny doses, help to reveal how nerve cells function — and potentially help to treat conditions from chronic pain to epilepsy.

Currently, most neuroscientists obtain their cone snail toxins from dead animals taken from the wild. But Bingham, a biochemist at Clarkson University in upstate New York, believes that the future lies with cone snail farming. Not only might it help conserve wild populations, he says, but it can also yield a wider range of useful toxins.

'Milking' the live snails is a hazardous business. One false move and Bingham could be dead in half an hour. Using forceps, he dangles a dead goldfish, the same length as Big Bertha, in front of her. Behind the bait, the condom is stretched over the mouth of the plastic tube.

"We tried all sorts of membranes, including sausage skin," says Bingham. "But for this species, condoms are the right thickness — and the snails don't like the lubricated ones." Bertha extends her proboscis and, with an abrupt flick, stabs the fish and the condom. She discharges a few microlitres of the toxin into the tube. That's enough to kill ten people.

Violent reaction

There are about 30 recorded instances of people being killed by cone snails — the molluscs are aggressive if provoked and can penetrate wetsuits with their sharp poison-loaded harpoons, which look like transparent needles. Human victims seem to suffer little pain¹, because the venom contains an analgesic component.

Today, the venom's painkilling properties are just one facet of a burgeoning field of research and drug development. Australian scientists first separated cone snail venom into its constituent parts in 1977 (ref. 2). Unlike most venomous animals, which produce one or a few poisons, a single snail

can produce up to 100 individual toxins.

When it comes to working out what each component does, Baldomero 'Toto' Olivera, a Filipino based at the University of Utah in Salt Lake City, has led the way. Throughout the 1980s, Olivera's students injected venom into the central nervous system of mice and found that different components had different effects — some would send a mouse to sleep, others would make it shake, run in circles or swing its head back and forth. The researchers showed that the different venom components had specific targets in the nervous system — they blocked different ion channels in cell membranes, which control the transmission of impulses by nerves, or specific receptors for the neurotransmitter chemicals that transmit signals from cell to cell³.

One snail's poison...

One venom component, called MVIIA and classified as an ω -conotoxin, causes tremors in mice. Olivera's group has found that it blocks a specific type of calcium channel that has been implicated in chronic neuropathic pain, caused by damage to the nervous system⁴. The toxin has since been developed as a drug called ziconotide and is now in advanced clinical trials for patients with cancer and AIDS who are suffering from pain that cannot be relieved by opiates⁵.

Different groups of conotoxins — most known by various letters of the Greek alphabet — target other ion channels, such as those for sodium and potassium, or receptors and transporters for neurotransmitters such as glutamate, serotonin, neurotensin and noradrenaline^{6,7}. "This is a real pharmaceutical cornucopia waiting to be tapped," says Bruce Livett of the University of Melbourne in Australia, who hopes soon to put



Milking time: Jon-Paul Bingham entices his molluscan menagerie to yield a sample of toxins.

an α -conotoxin, known as Vc1.1, into clinical trials for patients with diabetes who are suffering from the neuropathic pain that is a complication of the disease. Another toxin, which blocks a receptor for a specific neurotransmitter, is being investigated as a treatment for epilepsy. The goal is to damp down the hyperactivity that causes epileptic seizures.

In addition to their value as drugs (see Table, below), neuroscientists can use conotoxins as ultraprecise tools to selectively disable particular ion channels, or to block the action of individual neurotransmitters.

Toxin	Class	Species	Condition	Stage in development	Company website
Vc1.1	α	<i>Conus victoriae</i>	Neuropathic pain	Preclinical	www.metabolic.com.au
CVID	ω	<i>C. catus</i>	Neuropathic pain	Phase II clinical trials	www.amrad.com.au
MVIIA	ω	<i>C. magus</i>	Cancer pain	Phase III	www.elan.com
MrIA/B	χ	<i>C. marmoreus</i>	Neuropathic pain	Preclinical	www.xenome.com
Contulakin-G	Contulakin	<i>C. geographus</i>	Chronic pain	Phase II	www.cognetix.com
Conantokin-G	Conantokin	<i>C. geographus</i>	Epilepsy	Preclinical	www.cognetix.com



In addition, researchers can use fluorescently labelled toxins to find the pattern of distribution of ion channels in cell membranes. “We can home in on them,” says Bingham.

Fulfilling this research potential depends on a steady supply of cone snails. Which is why, once a year, Bingham joins a gang of ‘coneheads’ — amateur enthusiasts who collect the snails’ shells, heading out at night on to the beaches of the tropical Pacific, weighed down with backpacks carrying the batteries used to power their flashlights. “It’s difficult to find the snails,” says Bingham. But jumping up and down in the sandy shallows can help draw the creatures to the surface.

Unlike his companions, Bingham is intent on keeping his quarries alive. Each time he finds one, he pops it into a football sock so the snail draws its body into its shell and keeps its harpoon out of harm’s way — cone snails don’t like sock material. Bingham then takes the creatures back to the United States in a plastic container — “like a bucket

you keep your beers in” — after filling out copious customs paperwork.

The animals are difficult to keep in captivity. Bingham’s cone snail farm currently consists of just one tank of snails from a single species, *Conus purpurascens*. Although it bears the label ‘Danger — venomous snails’, at first glance, the tank looks as if it contains nothing but dirty brown sand. But Bingham calls it “cone snail Hilton”. He sprinkles a little fish-flavoured water into the tank. “Where else do you get waited on hand and foot?” he asks. The buried molluscs smell the hint of a meal and emerge over the next couple of minutes, until all nine are visible.

Cream of the crop

Once milking time is over, the next job is to put the samples through a high-pressure liquid chromatography machine, which separates the toxins within the venom and produces a printout with peaks that

correspond to individual compounds.

The conotoxins have very similar amino-acid sequences, and most are fewer than 30 amino acids in length. Using techniques including mass spectrometry and sequence analysis, Bingham is trying to work out the exact differences between the toxins.

Bingham has also found that his milked toxins are subtly different from those that are extracted from the venom glands of dead snails. The toxins are processed by enzymes as they move down the venom duct, he believes.

Perhaps his most fascinating finding, not yet published, is that the profiles of toxins produced by an individual snail are not uniform over time. Instead, the snails produce different toxins at different times — possibly influenced by external conditions such as temperature. If these conditions could be understood and controlled, says Bingham, neuroscientists could reap more valuable material.

For Bingham, these findings illustrate the superiority of cone snail farming over the traditional approach of extracting the toxins from dead snails. Conservationists, meanwhile, are worried that collection of the animals to provide material for neuroscience research and drug development could have a detrimental effect on wild populations. In October last year, Eric Chivian, director of the Center for Health and the Global Environment at Harvard Medical School in Boston, wrote a letter to *Science*⁸ expressing concern about the overuse of cone snails for scientific research. “No one really knows the number of cone snails that are killed,” says Chivian.

Neuroscientists working in the field think that the problem isn’t as bad as Chivian fears, but agree that cone snail farming could help reduce the toll on wild populations. “The milking technology could have a nice impact in preserving biodiversity,” says Olivera.

But achieving real sustainability will require the animals to breed in the lab, and this is more difficult than anyone anticipated. Bingham sheepishly admits that he is unable even to verify whether Big Bertha is a female. And, so far, he has not observed his snails mating. In the wild, cone snail orgies allegedly occur, but no one knows what triggers these erotic extravaganzas. For now, sex in the cone snail Hilton remains the stuff of Bingham’s dreams. ■

Laura Nelson is an intern with Nature in London.

1. Fegan, D. & Andresen, D. *Lancet* **349**, 1672 (1997).
2. Spence, I., Gillissen, D., Gregson, R. P. & Quinn, R. J. *Life Sci.* **21**, 1759–1769 (1977).
3. Olivera, B. M. *et al. Science* **249**, 257–263 (1990).
4. McIntosh, M., Cruz, L. J., Hunkapiller, M. W., Gray, W. R. & Olivera, B. M. *Arch. Biochem. Biophys.* **218**, 329–334 (1982).
5. Staats, P. S. *J. Am. Med. Assoc.* **291**, 63–70 (2004).
6. Livett, B. G., Gaylor, K. R. & Khalil, Z. *Curr. Med. Chem.* **11**, 1711–1719 (2004).
7. Alonso, D., Khalil, Z., Satkunathan, N. & Livett, B. G. *Mini Rev. Med. Chem.* **3**, 785–787 (2003).
8. Chivian, E., Roberts, C. M. & Bernstein, A. S. *Science* **302**, 391 (2003).



Breaking up is hard to do

Single-hulled ships are being rushed to the scrapyard in the wake of oil spills such as that of the *Prestige*. But will breaking them up cause environmental havoc too? Duncan Graham-Rowe finds out.

In November 2002, the *Prestige* oil tanker hit rough weather off the coast of Spain. The ship was badly damaged, the crew air-lifted to safety, and the vessel was led out to sea. Six days later it was torn in two, spewing 77,000 tonnes of oil into the sea. Kilometres of coastline were closed to the public as the thick black liquid spread over beaches and wildlife, prompting fears that the environmental disaster could be even worse than the one in Alaska caused by the *Exxon Valdez* 13 years earlier.

It's not known exactly what caused the *Prestige* to crack. What is known is that it was 26 years old and was a single-hulled ship. Such vessels have only a single skin between their cargo and the sea, leaving a slight margin of safety against a spill in case of collision. At the time of the accident, more than 2,200 such ships, many of them old and rusting, had already been earmarked for the scrapyard.

Under regulations drawn up in April 2001 by the International Maritime Organization (IMO) and the European Union, all single-hulled tankers had to be retrofitted or replaced with ships containing two hulls by 2015 — some types as early as 2007. But in the wake of the *Prestige* accident, those deadlines were cut to 2010 and 2005, respectively.

But this victory for the environment will come at a cost. As the deadlines approach there is a concern that there are not enough ship-breaking yards to cope with the impending deluge of ships. Instead, most of the vessels will end up being scrapped on beaches in India, Bangladesh and China, by workers without access to the tools and equipment that make scrapping safe.

At the same time, some people — includ-



ing both ship owners and environmentalists — argue that this rush to decommission old hulls won't necessarily stop oil spills. A double hull comes with its own problems, says Tim Wilkins, environment manager with Intertanko, an association that represents independent tanker owners. These cost more to maintain because roughly twice the surface area needs to be kept free from corrosion, and gases can build up between hulls, creating the risk of explosion.

Ground rules

Perhaps more importantly, double hulls are designed to protect against low-velocity impacts, says Wilkins. Yet almost invariably oil spills are caused not by collisions but by running aground, such as happened to the *Exxon Valdez* and the *Erika* tanker — which ran aground off the coast of France in 1999. "If both these ships had been double-hulled this would not have prevented the accidents," he says.

No one is arguing that the old, single-hulled ships should simply be allowed to stay



Sinking feeling: the demise of the single-hulled tanker *Prestige* (top left) left miles of coastline contaminated with oil and led for calls for such vessels to be scrapped quickly. But that could see hundreds of ships recycled on the beaches of Asian countries such as India (above).

at sea. But the rush to decommission them will cause its own problems. "It's a good thing to make tankers safer," says Marietta Harjono, coordinator of Greenpeace's ship-breaking campaign. "But we should be aware if we're simply shifting the pollution."

Oil isn't the only potential pollutant onboard a tanker. Even an empty ship can pose a serious environmental threat. A typical tanker contains heavy metals such as lead, mercury, cadmium and zinc in the paint of its hull — much of it intended to kill barnacles and other life that may try to adhere to the ship's belly. Other hazardous compounds used as antifouling agents include tributyltin, which is toxic to nerve cells and can accumulate in the blood, liver, kidney and brain. Polychlorinated organic compounds, which have been linked to cancer and liver damage, can be found in the insulation of old electrical cables. And asbestos litters old ships as a fire retardant.

In the West, says Harjono, tough regulations ensure the safe disposal of such hazardous materials. Ideally, ships are scrapped

in regulated dry docks or controlled quayside facilities where any pollutants released by the dismantling process can be contained.

But in Asia, where 92% of all ship scrapping takes place, there are precious few of these yards. Three years ago, even before the *Prestige* sank, Philippe Poirier d'Orsay, the now retired president of the Baltic and International Maritime Council (BIMCO) — a global shipping association — called upon China, where most of Asia's dry docks are located, to increase its ship-scrapping capacity to help meet the IMO deadlines. But not much seems to have been done.

On the scrapheap

The reality is that the vast majority of ships are going to be scrapped on beaches in India, China and Bangladesh, says Torben Strand, a senior manager at BIMCO. And it can be difficult to monitor activity there, even where legislation exists.

When ships are scrapped on beaches, their hazardous materials are often just burned in open pyres. Harmful fumes are then released directly into the environment — and into the people breathing the air. Useful materials such as asbestos are often removed without any protective equipment and left to dry on the beach before being resold.

In theory, there are some ship-scrapping

facilities in Western countries that could handle some of the load, but there is a definite 'not in my backyard' attitude when it comes to ship breaking, says Harjono. Governments would far rather shuffle the problem to a place where the cost of breaking is cheaper and the pollution problems are farther away.

The US Congress recently appropriated \$16 million for ship scrapping and now has several yards equipped for breaking. But after a bidding war, the government recently sent four of its ex-Navy 'ghost ships' across the Atlantic to the Able UK ship-breaking yard in Hartlepool. Their arrival was greeted by a mass of protests against what was called a cocktail of hazardous chemicals in the ships, and environmental groups claimed that Able UK did not have the right permits to deal with the 'toxic waste'. Those ships are still sitting in dry-dock in Britain, waiting for permits to roll in. Another nine former US Navy ships scheduled to sail for Britain seem destined to be broken up elsewhere — along with nearly 200 others also waiting to be scrapped.

Waste transfer

Environmental groups argue that shipping such waste to another country breaches the Basel Convention on the export of hazardous waste and so should be prevented, forcing Western countries to clean up their own mess. "It doesn't matter if you export a barrel of asbestos, or transport it as part of a ship's cabin," says Harjono. "The effect is the same." But many ship companies argue that the Basel Convention applies only to the transport of waste and so does not extend to the ships themselves.

Some countries are beginning to take steps that indicate they are aware of their global responsibilities. A European consortium is planning to build a new scrapping facility in the Netherlands to try to cope with the influx of ships, for example. A new type of environmentally friendly ship-recycling yard called an EcoDock is being planned at Eemshaven. Touted as a 'zero pollution' facility, this €45-million (US\$54-million) yard aims to use about one fifth of the workforce of a typical Chinese yard to scrap ships more safely and much faster than is currently possible.

Pioneered by a non-profit organization called the Stichting Tanker Ontmanteling Platform (STOP), the EcoDock should be able to use heavy cutting machinery to scrap a ship in just 23 days, says STOP chairman Doebren Mulder. In other yards the process usually takes 13 weeks and on the beaches it can take as long as eight months, he says.

If successful, STOP hopes to initiate up to 40 more yards based on this model, and has already begun negotiations with a number of governments to do so. Many of the governments in these talks are again in Asia, where

ship scrapping is big business and the raw materials — such as scrap steel — are in high demand. Giving Asia a greater capacity to scrap ships safely should be good news. Nevertheless, the first EcoDock won't be complete until July 2006, making it questionable how much of an impact these yards will have on the accelerated decommissioning process.

Across all the ship-breaking facilities available worldwide now, the industry can scrap about 28 million tonnes of steel a year, according to the London-based shipbrokers E. A. Gibson. Technically that is enough to scrap the 175 million tonnes of ships that need to be disposed of by 2010 — if all the vessels are scrapped at a constant rate. In reality, experts say, that is unlikely to happen.

Companies are reluctant to decommission their tankers ahead of the IMO deadlines because the ships are still earning their keep. So shipyards are currently running below capacity, looking for more ships to break, says Wilkins. As the deadlines approach the concern is that there will be a rush to scrap the remaining ships, overburdening the dry docks and forcing most ships onto the seemingly endless stretches of Asian beaches willing to take in more business.

Dirty work

In Bangladesh alone, more than 100,000 migrant workers are thought to depend on this industry. These are often vulnerable people with few other work options, who work for little more than a US dollar a day in some of the most dangerous conditions imaginable. According to Paul Bailey of the United Nations' International Labour Organization (ILO), ship scrapping is now considered one of the world's most dangerous occupations.

Figures taken from industrialized countries indicate that the rate of accidents and disease in ship breaking is very high compared with other industries. "We can assume that they are as bad, but probably even much worse, in developing countries," says Bailey. Labourers working on beaches receive little training and lack protective clothing or equipment, exposing them to hazardous materials and physical dangers including suffocation, falling debris, fire, explosions and electrocution.

In theory, there are provisions in place to help force this industry to a halt: beach scrapping is set to be phased out by 2012, according to guidelines set up in 2002 under the Basel Convention. But with the industry bringing in so much cash and raw material, it is not clear how this will be achieved.

In the meantime, some governments have tried to deal with the situation by making beach scrapping safer. India, for example, now refuses to allow ships to be scrapped on



Metal fatigue: an Indonesian woman collects rust during the beach scrapping of a ship.

its shores unless they have been certified as gas-free — meaning that the vessel has been checked for any potentially explosive pockets of gas. This has led to a drop in the number of explosion-related accidents, but has also lost India a lot of business as ship owners have taken their vessels elsewhere.

The ILO and the IMO have also produced guidelines for recycling ships safely. The IMO, for example, has introduced a Green Passport, which involves keeping an inventory of hazardous materials throughout a ship's life. Then, at least, ship breakers will know what they are dealing with when they come to tackle the vessel. The ILO guidelines

advise on the use of protective equipment when dealing with hazardous materials. But these guidelines are voluntary and the passport applies only to new ships.

If the rules and regulations could be made universal across the handful of countries involved in ship scrapping, says Bailey, then they would be much easier to enforce.

"What we want is a level playing field," he says. Alternatively, Western governments could introduce guidelines ensuring that ship owners can only sell their ships to scrapping companies that abide by ILO and IMO guidelines, says Bailey. One solution would be for ship owners to be made responsible for the conditions under which their ships are scrapped, he says.

In the absence of such action, some companies are taking matters into their own hands. The Dutch firm P&O Nedlloyd already has a policy of only allowing its ships to be scrapped in two dry-dock scrapping facilities in China, both of which abide by ILO guidelines. The company also precleans its ships before handing them over, removing any hazardous materials and providing protective equipment for workers. So far, 19 ships have been scrapped this way at an additional cost of no more than \$100 per tonne, says Tom Peter Blankestijn, P&O Nedlloyd's maritime-policy manager. But the company doesn't have any oil tankers, so this only applies to smaller ships.

Clean and dry

It is possible to scrap ships in an environmentally friendly way, as EcoDock and P&O Nedlloyd are showing, but it is not the cheapest option. "Some parts of the industry are very much in favour of changing the way ships are scrapped," says Frank Stuer-Lauridsen, chief environmental project manager at Danish engineering consultancy firm COWI, which is investigating the impacts of accelerated ship decommissioning. But, he says, whether a large part of the industry is willing to invest the money to make it happen is another matter.

It is very easy to blame ship owners for taking advantage of cheap scrapping facilities, says Strand, but finding out which ones are clean and safe can be difficult. "We are trying to accommodate concerns about the environment," he says, but the onus should not rest solely on ship owners. "If the ILO wants ship owners to sell to responsible yards then the ILO should supply a list of them," says Strand. This is on the cards, says Bailey.

In the meantime, as the ships queue up to be torn to bits and the deadlines loom, it is worth remembering that the motivation for this rapid response is based largely on politics rather than scientific evidence or environmental concerns, says Stuer-Lauridsen. "It's the high-profile sinking of ships that has driven this," he says.

In the end, the change to double-hulled ships won't be a panacea for the environment, says Harjono. The situation on the beaches where these boats are scrapped is already bad — and is now bound to get worse. The number of ships being ripped apart on the sand has increased by about 25% each year for the past three years, she claims. The people carrying out the work may depend on the industry and they may be doing the world a service, but they are paying for it dearly. "They are paying with their health, they are paying with their environment and they are paying with their lives. We think that is unacceptable," she says. ■

Duncan Graham-Rowe is a freelance writer based in Brighton, UK.

Rice, research and real life in the field

In the spirit of science, we should ask why studies don't reflect farmers' experiences.

Sir—I read the News Feature “Feast or famine?” (*Nature* 428, 360–361; 2004) with great interest, because I have been responsible for introducing the System of Rice Intensification (SRI) in the Indian state of Andhra Pradesh since the summer season of 2003. The experiences of farmers are quite different from what is reported by sceptical scientists.

SRI results are not “a miracle”; they are quite explainable. Planting young seedlings carefully and at wider spacing gives the plant more time and space for tillering and root growth. Careful water management, keeping the field wet and not flooded, gives better yield because it supports healthy root growth. This practice should be encouraged everywhere, as the whole world is facing water shortages. Weeding rice fields with a rotary weeder helps by churning the soil and incorporating the

weed biomass as it aerates the root zone. This encourages microorganisms to proliferate and promotes healthy soil. All these practices are known to agronomists — there is nothing new or magic about them.

The costs of SRI are low and its potential productivity is very high, which is more important than ever now that the Green Revolution technologies are showing signs of fatigue.

Even when the Andhra Pradesh farmers were unable to implement all aspects of SRI the first season, just planting young seedlings carefully at wider spacing with somewhat better water management resulted in yield increases of more than 2 tonnes per hectare compared with conventional methods using higher inputs. In 167 on-farm trials, the average yield obtained using SRI practices was

8.1 tonnes per hectare, compared with 5.67 using conventional practices. Average productivity in Andhra Pradesh is 3.89 tonnes per hectare. With more experience, still higher yields may be possible.

Rice yields all over the world have levelled out under the present system of flooded cultivation. We need to be looking for alternatives to existing practices, with an open mind. SRI is still evolving and I hope that the scientific community will collaborate in refining the technology and working out the scientific reasons for the reported higher productivity. This would be more constructive and more in the spirit of science than dismissing SRI with limited data and preconceptions, when the experiences of farmers are so positive.

A. Satyanarayana

ANGR Agricultural University, Rajendranagar,
Hyderabad 500 030, Andhra Pradesh, India

Rice: location is vital in crop management

Sir—Your News Feature about rice cultivation, “Feast or famine?” (*Nature* 428, 360–361; 2004), is a classic example of how the debate on a potentially interesting technique can be blurred by its opponents and proponents. In this case, proponents of the System of Rice Intensification (SRI) are reported to claim miracle yields of 15 tonnes or more per hectare, which opponents dispute. However, both decline to consider the location-specific conditions under which, for example, low yields in Madagascar (2 tonnes per hectare) were doubled or even tripled using SRI; these are not miracle yields but still considerable for poor farmers in marginal areas.

There is evidence that such an increase is closely related to better water management, for example non-flooded conditions during the vegetative growth phase of rice in Madagascar's highly reducing soils (J. F. Vizier *et al.* *Agron. Trop.* 45, 171–177; 1990). Iron toxicity is probably responsible for depressed yields under permanently flooded conditions in these soils. This could be a starting point in the better understanding of SRI. A closer look at the world's rice-growing areas could also reveal similar acidic soils that would potentially benefit from improved water management.

Even a moderate increase in rice yields (2–4 tonnes per hectare) as a consequence of reduced flooding cannot be maintained unless other crop-management practices are changed as well. For example, nutrient

inputs would need to be adjusted to account for the removal of more crop nutrients with the harvest, and weed control will require more attention under non-flooded conditions.

Other factors, such as decreasing farm labour in many rural areas and increasing water scarcity, pose important challenges to rice ecosystems. Only when the entire crop management is geared towards these often location-specific conditions, may new management practices — SRI could be one of many — contribute to the United Nations' goal of reducing hunger and poverty by half in a sustainable way.

Huib Hengsdijk, Prem Bindran

Plant Research International, PO Box 16,
6700 AA Wageningen, The Netherlands

Opening the chamber of peer-review secrets

Sir—I follow with interest the ongoing correspondence about peer review (*Nature* 427, 196 & 428, 255; 2004) and agree with the importance of giving recognition to quality reviewers. One option that has not had much attention is making the peer-review process public, so that the whole scientific community can benefit.

For example, *Neurosurgery* will print a short comment by reviewers highlighting why they felt that the paper deserved publication and what makes it new. One way to extend this idea would be to make each review, and the authors' responses to

the reviews for every article, available to all readers. This documentation could be posted as supplementary material on the Internet. Reviewers would have the option of remaining anonymous if they wish.

The educational nature of this information would be invaluable for many reasons. To cite a few: young researchers could learn how to publish outstanding papers and address criticisms; readers could be made aware of the limitations of certain approaches; and we would have a historical record of how peer review improves research findings.

Another practice, which I have found quite useful when reviewing for *Neoplasia*, is online dialogue between reviewers and authors. Reviewers' comments are posted anonymously online and authors can respond and clarify. The reviewer can help authors improve their studies by explaining why certain experiments were weak. The authors can clarify to the reviewer some aspects that may have been misunderstood in the review, possibly influencing the editorial decision.

This dialogue is unlimited and not mediated by an editor. At the moment it happens in a restricted area of the website, but it could be posted online with the accepted article. Once again, it would be the reviewer's decision whether or not to reveal their identity and assume responsibility for their review.

Erwin G. Van Meir

Departments of Neurosurgery, Hematology and
Oncology and Winship Cancer Institute,
Emory University School of Medicine,
1365C Clifton Road, Atlanta, Georgia 30322, USA

More than meets the eye

Earth's real biodiversity is invisible, whether we like it or not.

Sean Nee

The world has creatures that 'breathe' iron and uranium, using these elements in the same way as we use oxygen. Others thrive in the equivalents of hot sulphuric acid or floor stripper, and others again live in solid rock. In terms of biomass, the single most abundant form of life on Earth is ocean-sediment-dwelling bacteria and, numerically, the most abundant life-forms in the ocean are viruses. We are still at the very beginning of a golden age of biodiversity discovery, driven largely by the advances in molecular biology and a new open-mindedness about where life might be found. But for this golden age to be as widely appreciated as it should, our view of the natural world must change — as radically as did our view of the cosmos when we began looking at it with technologies that allowed us to see more than can be seen with the naked eye.

For all of the marvels in biodiversity's new bestiary are invisible. As many pundits have pointed out (usually in the restricted context of bacterial diversity), by any criterion — biomass or numbers of individuals — life on Earth is microscopic. It is the new generation of explorers of this 'invisible' world who are transforming our world-view beyond recognition. Yet a six-article Insight special on biodiversity in this journal in May 2000 scarcely mentioned anything that could not be seen unaided.

This is not to say that the papers were not a fair reflection of what biodiversity science is at the moment — that is, firmly fixated on the visible world. And I am not suggesting that senior commentators on biodiversity, such as Robert M. May and Edward O. Wilson, have been blind to the invisible. But it is now time for biologists — by whom I mean people who think of themselves as biologists, zoologists, botanists and ecologists — to cease presenting to their students and the public a perspective of life on Earth that is so biased towards the visible. This will not be easy. The first part of the challenge is accepting that the contribution of visible life to biodiversity is very small indeed.

Types of diversity

Phylogenetic diversity. The old 'five kingdoms' view of the relationships between organisms gave prominence to macroscopic creatures, with three of the kingdoms being animals, plants and fungi. But even under this classification, the invisible world asserts itself. The phylogenetic rank just below kingdom is phylum, and 40% of animal



Mr Spock — unemotional role model.

phyla are all or partly microscopic (numerically, four out of five animals are microscopic nematodes). But a new view of life's interrelationships has emerged from molecular data, in particular from the DNA sequences of genes that encode important RNA molecules found in all organisms. On the tree of life (see opposite) based on analyses of this 'small-subunit ribosomal RNA' (ssRNA), visible life consists of barely noticeable twigs. This should not be surprising — invisible life had at least three billion years to diversify and explore evolutionary space before the 'visibles' arrived.

The branch lengths in this figure are informative about diversity in terms of degree of divergence, and not just in the RNA genes used to construct the tree. Creatures that are distantly related according to their RNA genes are genetically divergent in many other respects. Consider the microsporidia, classically placed as a long branch at the 'base' of the Eukarya, one of the three main domains of life. These are remarkable intracellular parasites; their spores are wrapped in what is essentially a flexible hypodermic needle, and they infect a cell by piercing it and injecting themselves. In humans they cause diarrhoea. They are almost certainly misplaced in this tree as a result of a well-known artefact called 'long branch attraction', whereby very long branches get grouped together.

Microsporidia are now thought to be highly divergent fungi — highly divergent indeed!

Metabolic diversity. Life requires energy. Visible life exploits only one of the many possible metabolisms to get energy, taking electrons from organic carbon compounds and giving them to oxygen — breathing oxygen with which to burn food. This is mainly how animals and plants fuel themselves, although plants, of course, use light to manufacture their own food. In comparison, the invisible microbial world has a far greater metabolic repertoire. As I have mentioned, there are creatures that 'breathe' metals, using them as electron acceptors; others use metals as electron donors — they 'burn' metals. A quarter of a century ago it was predicted, on energetic grounds, that a creature should exist that burns ammonia with nitrite for energy (who says biology is not a predictive science?). Such bacteria were recently discovered¹.

Visible life can use one inorganic energy source — light — to manufacture organic compounds, but invisible life can also exploit chemical energy. Chemical energy almost certainly fuelled the first creatures that lived on Earth, and microbes that exploit a large variety of such sources provide the basis for the rich communities that thrive in the pitch darkness around hydrothermal vents on the sea floor — which were unknown to us until a quarter of a century ago.

The major biogeochemical cycles on which we depend were in place three billion years ago, long before the appearance of visible life, and are today maintained by the 'invisibles' and their vast range of metabolisms. The persistence of life on Earth depends on the endless recycling of essential elements such as sulphur, iron and nitrogen through their various forms. Thanks to macroscopic life, ecosystems are more productive and nutrient cycling is more rapid; animals and plants are good at churning things up. But the 'visibles' have introduced no qualitatively new features to global ecology.

Environmental diversity. Our awareness of the astonishing range of environments in which microbial life can thrive continues to expand. The superheated, pressurized environment of an autoclave, used for sterilization at 120 °C, would be ideal for the growth of 'strain 121', an 'iron-breathing' creature from a hydrothermal vent, which survives at even 130 °C. The greatest extremes of acidity, alkalinity, salinity and so on are not so extreme as to preclude invisible life. The Dead Sea is very much alive — it just doesn't contain fish.

Although representatives of the bacteria and archaea are the most famous extremophiles, microbial eukaryotes that occupy this stage are now also being discovered. Life in the Rio Tinto (river of fire) in Spain — which is metal-contaminated and as acidic as your stomach — is dominated by an extremely diverse array of them. A combination of the same molecular techniques used to study bacterial diversity in nature, as well as a willingness to look for them in supposedly inhospitable places, is revolutionizing our understanding of eukaryotic diversity, just as it is doing for bacterial diversity.

Morphological diversity. Visible life wins hands down in this category, although the invisible world puts up a good show with diatoms. Colours, noise, outlandish behaviour, emotion and consciousness — these must count for a lot. But none of it is really necessary for life on Earth, unlike the activities of invisible organisms, which terraformed the planet for visible organisms, creating and maintaining the atmosphere and nutrient cycles on which we depend.

Even the oxygen content of the atmosphere depends entirely on the invisible world of phytoplankton. Although roughly half of the planet's oxygen is produced by terrestrial plants, this is mostly used up in terrestrial respiration. The primary reason why the oxygen content of the atmosphere does not dwindle to zero is because of the rain of dead and dying oceanic microbes from the surface waters to the ocean floor. Buried in the sediments, the oxygen they produced is not used up in their decomposition, resulting in a net gain of oxygen by the atmosphere and oceans. It is a measure of how recently we have really started to get to grips with the microbial world that the most abundant constituent of the plankton, *Prochlorococcus*, was discovered less than 20 years ago. For emotional reasons, we attach great significance to the morphological dimension of diversity — biologists are largely attracted to the subject in the first place by this emotion.

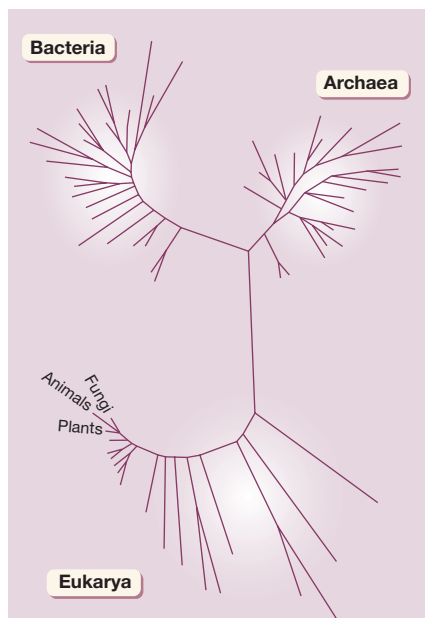
Departmental affiliations

A reflection of biologists' obsession with the visible world can be seen in the departmental affiliations of the explorers of the invisible world. With notable exceptions, these are rarely 'biology' or 'zoology'. Rather, more commonly, they are 'geology', 'biotechnology' or 'oceanography', and even 'civil engineering', 'planetary science' and 'astrobiology'. And astonishing new discoveries that challenge our views on the nature of life are more likely to be announced at a Geological Society meeting than a conference seemingly devoted to biodiversity.

Many of the explorers of the realm of the invisible are, of course, in departments of

microbiology. To update our view of life on Earth, biologists need to overcome the gulf between microbiology and biology that first arose about 150 years ago, and which retains much of its original character. Charles Darwin voyaged the high seas, dining at the captain's table, before retiring to his country house to develop his various ideas. Louis Pasteur, on the other hand, was hired as a consultant to the alcohol industry and discovered organisms that can live without oxygen. He experimented with heat treatments to prolong product shelf-life and developed vaccines to prolong human life.

Darwin was a biologist, Pasteur a micro-



Biodiversity through a molecular lens. This scheme is based on ssRNA gene-sequence data, and shows the relationships of organisms in the three main domains of life — Bacteria, Archaea and Eukarya (creatures with cell(s) like our own). Visible organisms are found among the plants, animals and fungi. Yet not only are these groups just twigs on the tree of life, but many of their members are invisible as well.

biologist. Microbiologists are employed by the wine and sewage industries; biologists go to wine tastings and discuss biodiversity and conservation in between using the bathroom. In the index to the standard microbiology textbook (*Brock Biology of Microorganisms*) you will find entries such as 'jock itch', 'sewage fungus' and 'Swiss cheese' — these are not the sorts of topics that turn on biology students. One challenge for biologists is to find ways of making sewage fungus palatable to our students and the general public.

Microbiologists do not even give a 'biodiversity' label to their studies — they call them 'environmental microbiology'. A major journal publishing fundamental discoveries in this area is *Applied and Environmental Microbiology*. Most biologists interested in

biodiversity would not even recognize that this is a highly important journal for them. And even when discoveries are published in the high-profile literature, as they frequently are, biodiversity enthusiasts are unlikely to register such titles as "An archaeal iron-oxidizing extreme acidophile important in acid mine drainage"² — which reports the discovery of creatures that are surely at least as astonishing as "Striped rabbits in Southeast Asia"³. If biologists want to keep up with this new golden age of biodiversity study, then there is no choice but to learn or relearn our basic chemistry, redox reactions in particular.

Emotional bias

We are sentimental about visible nature yet utterly self-centred about the invisible — simply happy with it if it is useful, such as bakers' yeast, or bent on its extermination if it is unpleasant or harmful (for example, a yeast infection). And, if it is not obviously important to us in one way or another, we do not even give it a thought. As May has said⁴: "As one moves down the size spectrum of organisms, from the romantic large mammals and birds, through non-descript small arthropods, on down to protozoan, bacterial and viral species, not only does concern for diversity and conservation fall away, it even changes sign."

As biologists, we must recognize this emotional bias and try to control it, like Mr Spock. Honesty to ourselves in considering what life is — and in presenting an honest picture to students and the public — demands it. Consider 'syntrophy'. This is the dominant ecological relationship in worlds without oxygen⁵ — such as sediments and your gut — and is the phenomenon whereby the metabolic waste product of one species is the primary food or energy source of another (there are stricter definitions). The syntrophic networks that take in organic material provided by the oxygenated worlds and output the final waste products of carbon dioxide and water can be complex. But syntrophy is not mentioned in any major ecological textbook. Ecology students carry around inside their guts highly diverse ecologies, and this is simply never even mentioned. This neglect of the invisible world is no longer any more acceptable than, say, teaching astronomy but ignoring the existence of galaxies beyond the Milky Way, or teaching physics while refusing to discuss anything smaller than a pin head. ■

Sean Nee is in the Ashworth Laboratories, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK.

e-mail: sean.nee@ed.ac.uk

1. Strous, M. et al. *Nature* **400**, 446–449 (1999).
2. Edwards, K. J., Bond, P. L., Gihring, T. M. & Banfield, J. F. *Science* **287**, 1796–1799 (2000).
3. Surrage, A., Timmins, R. J., Hewitt, G. M. & Bell, D. J. *Nature* **400**, 726 (1999).
4. May, R. M. *Phil. Trans. R. Soc. Lond. B* **330**, 293–304 (1990).
5. Fenchel, T. & Finlay, B. J. *Ecology and Evolution in Anoxic Worlds* (Oxford Univ. Press, 1995).

Paying for the pills

How much does it really cost to put a new drug on the market?

The \$800 Million Pill: The Truth Behind the Cost of New Drugs

by Merrill Goozner

University of California Press: 2004. 297 pp.

\$24.95, £16.95

John E. Calfee

The \$800 Million Pill is open to trenchant criticism of some of its most basic points, but anyone who reads it will find some wonderful material. Merrill Goozner is a member of a small but influential group of journalists who, by training or diligence, have learned enough science and medicine to bridge the gap between science and policy without making fools of themselves. They perform a valuable service, not least because they can write well. Here Goozner effortlessly pulls the reader through some extremely complicated topics.

The title refers to a much cited and debated paper by Joseph DiMasi, Ronald Hansen and Henry Grabowski (*J. Health Econ.* **22**, 151–185; 2003). Based on extensive but proprietary industry data, this study concluded that, on average, each drug introduced into the market from 1990 to 2001 involved total costs of US\$802 million. This figure takes into account the opportunity costs of capital and the costs of the numerous failures that litter the path of drug development.

Relying mainly on non-scholarly critics such as Public Citizen (a lobby group in Washington DC founded by Ralph Nader) and the Global Alliance for TB Drug Development, Goozner argues that DiMasi and colleagues grossly overstated the true costs of developing useful drugs. He accepts instead the Global Alliance's estimate that discovering and developing a new tuberculosis drug would cost only US\$115–240 million. Essentially, Goozner argues that the editors and reviewers of the *Journal of Health Economics* overlooked some basic methodological flaws in the paper. His argument is unpersuasive, though, partly because it defies sound economic reasoning on opportunity costs. It also ignores important limitations in the Global Alliance's study of developing a hypothetical drug for a single condition (tuberculosis) and using relatively small clinical trial sizes, non-market cost data and highly conjectural failure rates.

But 90% of this book is about other matters. Goozner pursues two main themes and many lesser ones. First, he claims that partly because of industry obfuscation, the world little appreciates the crucial role in the creation of new drugs played by research funded by non-industry sources, such as the US National Institutes of Health (NIH). His case



Strong medicine: pharmaceutical companies take a huge gamble when they develop a drug.

studies focus on HIV/AIDS (roughly one-third of the book), cancer and the biotechnology drugs erythropoietin and imatinib mesylate (Gleevec).

The details here are fascinating and many of his points are well delivered. But they fail to carry his wider argument: that the NIH and other non-industry sources typically complete the most difficult part of developing essential drugs and then turn them over to a few private firms to market them at immense profit. This argument is hard to square with the book's own repeated examples of the NIH's difficulties in overcoming the "reluctance" of individual firms to pursue the costly clinical trials needed to move a promising drug towards approval by the US Food and Drug Administration (FDA). This reluctance is understandable. The supply of promising molecules discovered with NIH money is very large. Anyone spending their own money will be very selective in deciding which of these is worth an investment of tens of millions of dollars to find out whether it is the rare success rather than the latest addition to the long list of costly failures.

This is where the economics gets ahead of Goozner, despite his formidable dedication and skills. A crucial point is not just the uncertainties of drug development, but the fact that, in the private sector, firms have to eat their mistakes. When Merck saw no fewer than four new drugs fail in phase 3 trials in 2003, the costs were borne by Merck stockholders and employees; there was no prospect of Merck covering those losses by simply raising the price of its existing drugs.

Goozner's other broad theme is that much of the industry's research and development is wasted in the pursuit of 'me-too'

drugs that deliver easy profits but little medical benefit. Like most critics of follow-on drugs, Goozner does not offer a list of the unnecessary drugs that account for half of industry revenues. As recently pointed out by Thomas Lee (*New Engl. J. Med.* **350**, 211–212; 2004), follow-on drugs typically cater for patients who are poorly served by existing drugs, and provide a strong dose of price competition. Economists have documented the potency of these market forces.

Goozner's recommendations reflect his major themes. Eschewing government controls over prices or profits (mainly, it seems, for political reasons), he seeks better information — of the most expensive kind. He wants more publicly funded clinical trials comparing new and old drugs. More importantly, he wants the FDA's new drug approval standards to require comparative clinical trials, but this would greatly increase costs without yielding comparable benefits. The FDA has repeatedly pointed out that comparison trials have to be very large to achieve the statistical power necessary to reveal genuine therapeutic differences. If the FDA requires such trials, they will probably show that the typical new drug helps at least some patients more than existing therapies, and that it still meets the FDA's standard for a safe and effective treatment. This approach is more likely to increase the costs of drug development than streamline the system. ■

John E. Calfee is a resident scholar at the American Enterprise Institute, 1150 17th Street, NW, Washington DC 20036, USA.

John E. Calfee's employer receives financial contributions from individuals, foundations and corporations, including pharmaceutical firms. Some of his research received partial funding from the pharmaceutical industry.

D. MEYLER/CORBIS

Requiem for a supercollider

A Hole in Texas

by Herman Wouk

Little, Brown: 2004. 278 pp. \$25, Can\$37

Geoff Brumfiel

"Gentlemen, the Chinese have got the Higgs boson."

So begins *A Hole in Texas*, Herman Wouk's latest novel, which revolves, rather improbably, around the world of high-energy physics. The 88-year-old Pulitzer Prize-winning author of *War and Remembrance* and *The Winds of War* is famous for his depictions of wars and the men who fight them, but his twelfth novel concerns exotic particles and the men (and women) who discover them. The result is an amusing and at times arcane yarn that will make a good summer read for scientists — especially those who have had a frustrating time explaining their work to governments or the popular press.

Set in the not-too-distant future, the book has as its protagonist Guy Carpenter, a silvery, fifty-something physicist who once worked on the US Superconducting Super Collider (SSC), which really existed. As Wouk explains in the novel's foreword, the supercollider was a giant US particle accelerator designed to find the Higgs boson, an elusive particle that many physicists believe endows other particles with mass. After spending billions of dollars on the project, Congress withdrew support in 1993, leaving behind a couple of thousand unemployed physicists and a partly dug tunnel in Waxahachie, Texas — the 'hole' of Wouk's title.

Carpenter has moved on since the SSC was shut down, but like any good dime-novel hero, his past comes back to haunt him. When word gets out that the Chinese are about to publish the discovery of the Higgs boson (in *Nature*, of course), Carpenter is asked to explain the significance of the find to a movie-star-turned-Congresswoman. By the time he arrives in Washington, the media are swarming around the story, and a secret memo is circulating on Capitol Hill that the Chinese are building a 'boson bomb'. Word of the bomb gets out to the press and Carpenter is rapidly drawn into the ensuing national panic. He is pursued by a "pestiferous ferret" of a reporter, who is determined to find out more about Carpenter's graduate-school liaison with the female physicist now leading the Chinese programme. He is wined and dined by Hollywood bigwigs, who want to make a movie about the boson bomb. And he is called to testify by an unscrupulous Congressman who voted to kill the supercollider years ago and is now seeking a scapegoat on whom he can lay the blame.

If all this sounds a little absurd, that's the



Billion-dollar hole: Congress pulled the plug on the real Superconducting Super Collider in 1993.

whole point. Wouk's work is satire, and it delivers its fair share of laughs (the movie about the boson bomb is set to star Julia Roberts as "the astronaut"). At its heart is an idea summed up nicely by one of the book's physicists: "Science today needs huge funding, and we're funded by a scientifically illiterate society."

That message would be fine if it were not for the breathtaking arrogance of the book's physicist characters who deliver it. They go to great lengths to trumpet the ignorance of the media, Congress and the public, and one even goes so far as to say that the gap between someone who knows quantum mechanics and someone who does not is as wide as that between man and monkey. This too could be satire, but reading page after page of the stuff you get the uneasy feeling that Wouk might actually mean it. By the way, I got an A minus in quantum mechanics.

This is far from being the book's only flaw. The scientific explanations are pat and usually come in the form of long e-mails that bog down the plot. A trip to the hole in Waxahachie takes far longer than it should and reads like a giant I-told-you-so. The women of the novel, whether they be hard-nosed members of Congress or top-notch physicists, fall helplessly into Carpenter's powerful arms. And everybody's discussion of the Chinese people can, at times, verge on racism ("They're touchingly sentimental, the Chinese, once you're past those forbidding Asian features and the different eyes," Carpenter declares).

The book's ending also falls flat. Without spoiling too much, I'll say that the furore over the Chinese discovery forces the public to focus on the importance of science. In the

modern world that seems unlikely. After all, the Chinese launched their first man into space last autumn at a time when the US shuttle programme was out of commission, but no alarm bells sounded. The threat posed by low-tech extremists seems far scarier than the rise of Chinese science. It was that fact, along with the rather wistful view of the SSC, that left me feeling that *A Hole in Texas* is something of an anachronism. Still, it will make good beach reading for researchers who want a scientific thriller and a bit of a chuckle.

Geoff Brumfiel is *Nature's* Washington-based physical sciences correspondent.

Bright blue dot

The Privileged Planet: How Our Place in the Cosmos is Designed for Discovery

Guillermo Gonzalez & Jay W. Richards

Regnery: 2004. 464 pp. \$27.95

Douglas A. Vakoch

"Our planet is a lonely speck in the great enveloping cosmic dark," wrote astronomer Carl Sagan in *Pale Blue Dot* (Random House, 1994). Reflecting on the image of Earth sent back from the Voyager 1 spacecraft in 1990, he suggested that "our posturings, our imagined self-importance, the delusion that we have some privileged position in the Universe, are challenged by this point of pale light."

In *The Privileged Planet*, astronomer Guillermo Gonzalez and the philosopher and theologian Jay W. Richards present a markedly different view. They argue that our

Exhibition

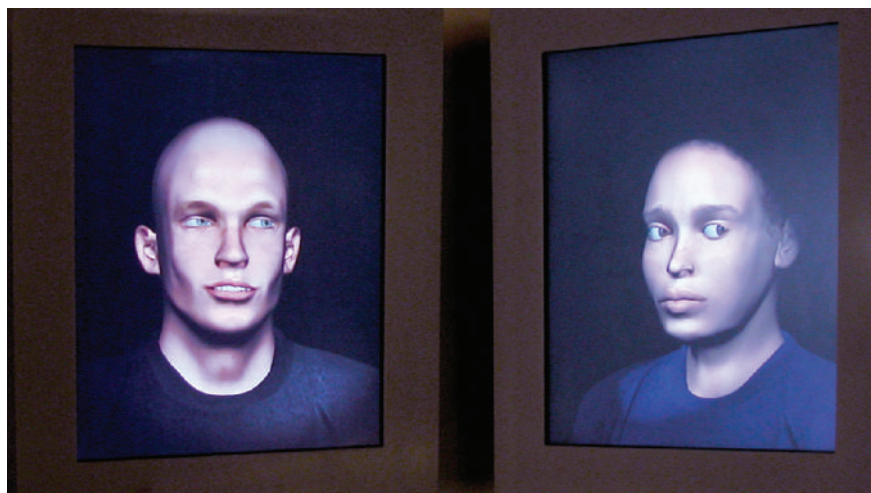
Eyes that follow you around the room

Since 1979, the industrial city of Linz in Austria has been the unlikely home of Ars Electronica, a centre for the development and dissemination of digital art and media. The centre has evolved to include an annual festival, a 'cyberarts' competition and 'The Futurelab', a media lab where interdisciplinary collaborators design and engineer new installations.

To celebrate Ars Electronica's 25th anniversary, New York City's American Museum of the Moving Image is hosting Digital Avant-Garde, a series of exhibitions and screenings of projects that featured prominently in earlier competitions. Linz will have its own anniversary celebration during the first week of September.

The New York series, which runs until 18 July, includes John Gerrard's *Portrait Diptych: Nadia* (formerly called *Networked Portrait*), an interactive portrait that can be changed with the touch of a finger on a screen. The installation, shown here, consists of two liquid-crystal-display touchscreens, each presenting the computer-generated image of a particular individual.

By dragging a finger across an eye or the



corner of the mouth in one of the images, the viewer can impart some emotion to the otherwise expressionless visage. When the screens are turned toward each other, the other face responds by subtly changing its own expression. The two faces then continue to respond to each other.

The portrait is seen as one of the last bastions of permanence at a time when images can be erased or modified at will. But even the portrait, it seems, can be digitized and updated continually — for example, to reflect one's mood.

Alan Packer

www.aec.at/en/index.asp

planet is significant, perhaps being uniquely situated to foster both complex life and scientific discovery. Gonzalez and Richards believe that intelligence may be a rare phenomenon in our galaxy. They make their case, in part, by drawing on arguments others have given before, such as Peter D. Ward and Donald Brownlee in *Rare Earth* (Springer, 2000). But they go far beyond estimating the prevalence of extraterrestrial life. As they summarize: "The myriad conditions that make a region habitable are also the ones that make the best overall places for discovering the universe in its smallest and largest expressions. This is the central argument, the central wonder of this book."

Drawing on a framework for inferring design proposed by philosopher and mathematician William Dembski in *The Design Inference* (Cambridge University Press, 1998), Gonzalez and Richards argue that the correlation between the conditions that make habitability possible and those that make it possible to learn about the Universe is so exquisite and improbable as to suggest intelligent design.

For example, by raising the tides the Moon helps move nutrients from the land to the ocean, fostering life in the intertidal zone. If the Moon were farther from the Earth, the authors argue, it would need to be much larger to have the same effect on the tides; if it were closer, it might well be less spherical, causing other problems. So, Gonzalez and Richards argue, the Moon is located at a distance from Earth that is very conducive to life. It is also at

an optimal distance for scientific discovery (or measurability), they contend, as the disk of the Moon provides a perfect eclipse of the Sun. They argue that without such perfect solar eclipses, humans would have been deprived of important information about the Sun's chromosphere.

A single such instance linking habitability and measurability might be dismissed as a coincidence, but Gonzalez and Richards claim that the correlation across many phenomena cannot be explained by chance. By including astronomical and geological phenomena in their search for evidence of purpose in the cosmos, *The Privileged Planet* expands the discussion of intelligent design far beyond its usual contemporary focus on the complexity of biological systems and the fundamental physical constants.

Ultimately, however, the authors are in a poor position to argue that Earth is optimally located for both habitability and measurability. They try to establish habitability requirements by comparing Earth with other locations in the galaxy. Unfortunately, we lack the data required for a well-reasoned comparison. If we had many examples of planets that do and do not bear life, and an explanation for why the conditions on some planets led to life while those on others did not, we might be able to establish an accurate metric of habitability. Until then, we are forced to extrapolate measures of habitability from a sample of one inhabited planet.

Regrettably, any judgements of optimality may be biased by the local conditions and

historical contingencies through which our life and our science have evolved, rather than accurately reflecting the range of possible preconditions for habitability and measurability. Potential readers of *The Privileged Planet* would do well first to familiarize themselves with the biases that can result from this kind of selective sampling. A good primer is Nick Bostrom's *Anthropic Bias* (Routledge, 2002).

Caution seems especially in order given that the authors have intentionally limited themselves to our knowledge of the Universe gained through selected observational sciences, such as comparative planetary geology, solar physics and astronomy, rather than including more laboratory-based sciences. Similarly, although the authors attempted to avoid cherry picking instances of measurability that support their position by focusing on important observations in these fields, the vagueness of such a criterion makes their selection rather subjective.

So far, Earth is the only planet we know that has the privilege of bearing life that searches for signs of other intelligence — whether in the form of other technological beings transmitting evidence of their existence or through patterns indicating underlying design. It may be some time, however, before we can accurately judge whether our blue dot is — as planets go — commonplace, unique or somewhere in between. ■

Douglas A. Vakoch is at the SETI Institute, 2035 Landings Drive, Mountain View, California 94043, USA.

Mentors and manipulation

How Nobel assistance helped a young researcher test a crazy idea.

G. J. V. Nossal

Frank Macfarlane Burnet and Joshua Lederberg, Nobel laureates of 1960 and 1958, are each giants: Burnet is the grand old man of Australian virology and immunology; Lederberg is the brilliant founder of the new discipline of bacterial genetics. During three heady months in 1957, I was the meat in their sandwich, and had the good fortune of instigating a turning point in the exciting problem of how cells make antibodies.

Mammals can make antibodies to virtually anything, including synthetic chemicals. How can the cells of the immune system, the lymphocytes, fabricate such diversity? Paul Ehrlich was the first to offer an answer in 1900 with his 'side-chain' theory. He imagined stereochemical configurations on antibodies that were complementary to corresponding configurations on foreign agents (antigens). He postulated that antibodies existed as 'side chains' on the surface of cells. When an antigen came along, it united with the appropriate side chains, which the cell then overproduced and — hey presto! — you have antibodies.

Kari Landsteiner doubted this, because tiny changes in a synthetic antigen altered the specificity of antibodies. Given the number of possible chemical antigens, it beggared belief that there could be such a variety of pre-existent side chains. Ehrlich's views gave way to the 'direct template' theory, which suggested that antibodies moulded themselves directly against the antigen concerned. Niels Jerne challenged this 'instructionist' view by renewing the suggestion that antibodies were preformed, arguing that even if there were as many as 10^{11} different natural antibodies, there could still be a million of each type in a millilitre of serum. All the antigen had to do was to find the 'right' antibody and somehow catalyse its mass production. David Talmage was the first to suggest that Jerne's theory might work if the relevant antibodies were receptors on lymphocytes — and so the argument went back to Ehrlich! Burnet had independently come to the same conclusion and fleshed out what he termed the 'clonal selection' theory. If there were some random generator of diversity among antibody genes, switched on during the development of lymphocytes, each cell might end up with just one kind of antibody on its surface. When an antigen entered the body, it would find the right lymphocyte and stimulate it to divide and mass-produce antibody. The lymphocyte population could be seen as a repertoire of specificities.



Ideas in immunology: Frank Macfarlane Burnet.

I was a naive research fellow joining Burnet's laboratory fresh out of medical school, and was asked what I thought. Frankly, it all seemed a bit crazy. Not having the courage to say that, I ventured to point out a way of disproving clonal selection. If a single cell from an animal that was immunized simultaneously with two antigens could produce both antibodies, the theory was wrong. Moreover, I knew that Lederberg — soon expected to arrive at our laboratory as a Fulbright visiting professor — was expert in micromanipulation. Perhaps he would help me to get single lymphocytes into microdroplets for short-term tissue culture.

When Lederberg arrived, he was deeply fascinated by Burnet's theory, preferring selective to instructive processes. We started our brief collaboration and chose immobilization of motile bacteria by anti-flagellar antibody as our assay. We immunized rats with killed *Salmonella typhi* and *Salmonella adelaide*.

Our only microscope was an old-fashioned Bausch and Lomb, with vertical rather than inclined eye-pieces — a good way of getting a crick in the neck. To see swimming bacteria we needed dark-field microscopy but, alas, we had no dark-field condenser. Undaunted, I simply removed the condenser and replaced it with a large Australian penny placed in such a way that an arc of light fell on to the specimen micromanipulation chamber, resulting in a fair imitation of dark-field illumination. We borrowed an ancient micromanipulator from the University of Melbourne across the road, and away we went.

Lymph nodes from immunized rats were teased into a suspension of single cells, then

microdroplets that were hanging from a coverslip and surrounded by mineral oil were incubated for four hours. After this, 5–10 motile bacteria, first from one strain, then from the other, were micropipetted into each droplet. Regrettably, Lederberg had to leave before the first results trickled in, but our *Nature* paper (181, 1419–1420, 1958), although technically far from perfect, supported the one cell: one antibody concept.

Over four years, two in collaboration with Olavi Mäkelä, we studied 3,628 active single cells. Only two actually secreted antibody reactive with two bacterial strains, presumably representing cross-reactivities not detectable at the whole-serum level or, alternatively, binucleate cells, which are not all that uncommon among antibody-forming cells.

On his return to Finland, Mäkelä showed that each cell had its own unique fingerprint of antibody specificity, a great prelude to the qualities we now know monoclonal antibodies to possess, and in exact agreement with clonal selection. In a further collaboration with Gordon Ada, we made autoradiographs of single, antibody-forming cells from rats that had received a very highly radioactive antigen. In a situation where one silver grain above background would have represented four molecules of antigen, the mean grain count over 216 single antibody-producing cells was — 0.4. There simply was no antigen in them, so the direct template theory was dead.

But what was really needed for formal proof was to take single cells from unimmunized animals that had been fractionated to react with a specific antigen; then clone them *in vitro* and demonstrate that the antibody formed had the same specificity as the receptors on the original B cell. When Beverley Pike and I finally did this in 1976, the results occasioned little comment because, essentially, clonal selection had won the day. The generator of diversity of antibodies turned out to be a unique somatic mini-gene shuffling and reassembly process.

For six years after the turning point, my colleagues and I had the study of antibody formation by single cells to ourselves, as the micromanipulation techniques were far too tedious for most investigators. Then a much simpler plaque technique came along and everybody could play the game. By that time, we had discovered a fascinating new cell, the follicular dendritic cell, that held the key to immunological memory. But that turning point will have to wait for another day. ■

G. J. V. Nossal is in the Department of Pathology, the University of Melbourne, Victoria 3010, Australia.

Diversity in the deep blue sea

Peter J. Morin and Jeremy W. Fox

A large-scale survey of the diversity and abundance of plankton in different marine environments around the world has produced some thought-provoking similarities and contrasts with other ecosystems.

Much of what we know about how the diversity of life varies across environments comes from studies of large, 'charismatic' terrestrial organisms that typically attract the attention of ecologists¹. These studies show that diversity often peaks at intermediate levels of productivity, where productivity describes the rate of energy capture and its transformation into biomass by organisms. Little is known about whether similar patterns of diversity and productivity hold for the much smaller organisms (Fig. 1) that predominate in the world's oceans, and there are suggestions that some ecological patterns of single-celled organisms might differ in important ways from those of larger ones².

On page 863 of this issue³, Irigoien *et al.* report that the algae — phytoplankton — supporting food webs in the oceans, Earth's largest ecological realm, exhibit a unimodal diversity–productivity pattern similar to that described for many other systems. Revealing the unimodal pattern required the compilation and analysis of a database of algal species composition and biomass (where biomass serves as a reasonable surrogate for productivity) in more than 350 samples collected from oceans around the world. Although several other diversity–productivity patterns exist¹, the unimodal pattern appears repeatedly for a variety of organisms in different circumstances. In this respect, ecology exhibits some important generalities, which can now be extended to the oceans. The analysis also emphasizes the value of the new field of 'ecoinformatics' in uncovering patterns that emerge only from extensive data sets collected over large temporal and spatial scales.

Several factors could make phytoplankton diversity reach a peak at intermediate levels of productivity. Community history⁴, spatial niche differentiation⁵ and competition for multiple resources⁶ can all produce unimodal patterns. Community history refers to the stochastic, sequential arrival of species at a local site from a surrounding regional pool of potential community members. The arrival sequence can affect diversity (for instance, weaker competitors might persist if they arrive and become established before other species), and can interact with productivity to produce a variety of diversity–productivity relationships, including a unimodal pattern⁴. Spatial niche differentiation occurs when local sites contain different

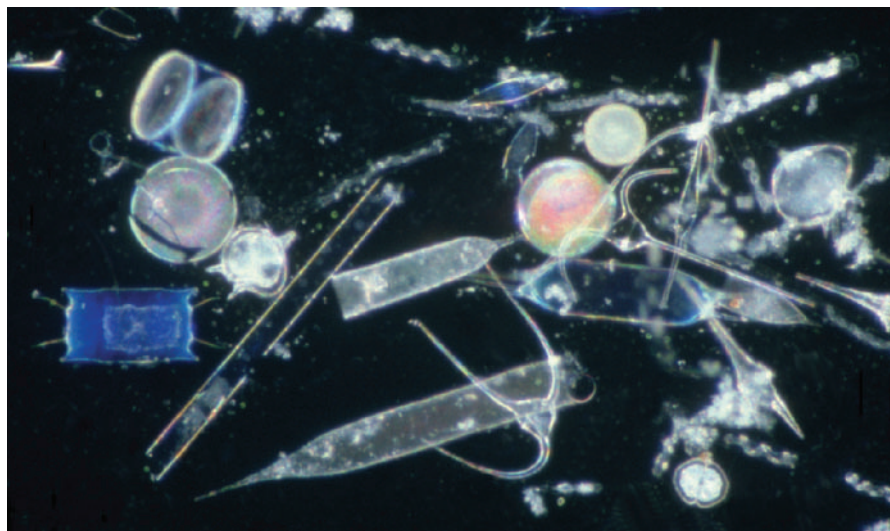


Figure 1 Plankton: little organisms that pose big ecological questions.

microhabitats (spatial niches), such as a lake surface or water column, to which species are differentially adapted, and which vary in their ability to support reproduction. Diversity is maximized in sites that are sufficiently productive to allow reproduction in all microhabitats but are not so productive that the species adapted to the best microhabitat produces sufficient offspring to 'swamp' other species⁵. The hypothesis of competition for multiple resources⁶ suggests that shading within dense algal blooms probably increases with productivity, so that phytoplankton at high-productivity sites compete more strongly for light than for nutrients. Superior competitors for nutrients or light would respectively dominate low- and high-productivity sites, whereas sites of intermediate productivity would support a diverse mixture of phytoplankton.

Shifts in the relative importance of competition and predation along productivity gradients⁷, in which superior competitors predominate at the low end of the productivity gradient and species that are well defended against consumers predominate in the most productive sites, can also create unimodal diversity–productivity patterns. In unproductive sites such as nutrient-poor areas of the open ocean, predators should be food-limited and rare, and only a few species of competitively superior phytoplankton will predominate. Highly productive sites, for example nutrient-rich coastal upwellings, should support abundant predators, which

then remove all but the most predator-resistant species from the phytoplankton. At intermediate sites, a mix of competition and predation should allow the coexistence of a maximal diversity of phytoplankton species.

Increases in average cell size within the phytoplankton assemblage with productivity provide some indirect support for the predation hypothesis. Work in freshwater systems shows that phytoplankton cell size mediates a trade-off between competitive ability and predation resistance⁸. The high surface-area-to-volume ratios of small cells should confer superior competitive ability for dissolved nutrients, whereas larger cells should be less vulnerable to consumption by microzooplankton, such as alga-feeding protists that are restricted to feeding on smaller cells. Dominance of large phytoplankton at high productivities is not predicted by alternatives to the predation hypothesis. For instance, a shift from competition for nutrients at less productive sites to competition for light at more productive sites⁶ probably does not drive shifts in phytoplankton diversity and cell size. Large phytoplankton compete poorly for light because of self-shading⁹, and so would not be expected to dominate highly productive sites if competition for light determined phytoplankton composition.

The situation becomes even more interesting for the zooplankton that consume algae. Irigoien *et al.*³ show that phytoplankton diversity is a poor predictor of zooplankton diversity. This contrasts with patterns

seen in many terrestrial systems, in which the greatest diversity of consumers occurs in association with the greatest diversity of primary producers. The lack of connection between phytoplankton and zooplankton diversity might result because the unicellular nature of the phytoplankton eliminates a major source of diversity for consumers. Unlike the spatial and structural complexity produced by, say, a canopy of tropical trees, the phytoplankton contribute little structure to their environment. Theory predicts that the morphological complexity of large terrestrial plants provides niches for the smaller organisms that exploit them, and the diversity of these organisms increases fractally as their own size declines¹⁰. By increasing the uncertainty of species associations in time and space, small size and hydrodynamic complexity might also reduce the frequency of co-evolved feeding relations that foster the correlated diversity of terrestrial plants and arthropods (herbivorous insects, pollinators and so on). In this sense, differences in the relative sizes of primary producers and their consumers in aquatic and terrestrial environments might contribute to fundamental differences in the ways in which these communities are organized¹¹.

What, then, determines zooplankton diversity? Irigoien *et al.*³ suggest that, as with phytoplankton, it stems from a shifting balance between competition for food and resistance to predators. But there are other possible explanations. Ecological theory has yet to thoroughly consider the mechanisms that might couple or decouple predator and prey diversity along productivity gradients, and this remains a promising area for research. As this paper³ shows, broad, cross-system comparisons, interpreted in the light of general theories, can reveal surprising commonalities in the diversity of life.

Peter J. Morin is in the Department of Ecology, Evolution and Natural Resources, 14 College Farm Road, Rutgers University, New Brunswick, New Jersey 08901, USA.

e-mail: pjmorin@rci.rutgers.edu

Jeremy W. Fox is at the NERC Centre for Population Biology, Imperial College London, Silwood Park Campus, Ascot, Berkshire SL5 7PY, UK.

e-mail: jeremy.fox@imperial.ac.uk

1. Waide, R. B. *et al.* *Annu. Rev. Ecol. Syst.* **30**, 257–301 (1999).
2. Finlay, B. J. *Science* **296**, 1061–1063 (2002).
3. Irigoien, X., Huisman, J. & Harris, R. P. *Nature* **429**, 863–867 (2004).
4. Fukami, T. & Morin, P. J. *Nature* **424**, 423–426 (2003).
5. Kassen, R., Buckling, A., Bell, G. & Rainey, P. B. *Nature* **406**, 508–512 (2000).
6. Tilman, D. & Pacala, S. in *Species Diversity in Ecological Communities: Historical and Geographical Perspectives* (eds Ricklefs, R. E. & Schluter, D.) 13–25 (Univ. Chicago Press, 1993).
7. Leibold, M. A. *Am. Nat.* **147**, 784–812 (1996).
8. Agrawal, A. A. *Acta Oecol.* **19**, 331–337 (1998).
9. Malone, T. C. in *The Physiological Ecology of Phytoplankton* (ed. Morris, I.) 433–464 (Univ. California Press, 1980).
10. Morse, D. R., Lawton, J. H., Dodson, M. M. & Williamson, M. H. *Nature* **314**, 731–733 (1985).
11. Hairston, N. G. Jr & Hairston, N. G. Sr *Am. Nat.* **142**, 379–411 (1993).

Planetary science

How Mercury got its spin

Stanley F. Dermott

The orbital period of Mercury and its period of rotation are known to be in a 3/2 ratio, but the chances of the planet reaching this state seemed so small as to be unfeasible — until now.

Like most of the large satellites in the Solar System, the Moon's orbital period and its period of rotation are the same: the Moon completes both an orbit of the Earth and a rotation about its own axis in 27.3 days, and hence always keeps the same face towards the Earth. But, in 1965, observations¹ of Mercury turned up a great surprise: the rotational period of that planet is only two-thirds of its orbital period (59 days compared with 88 days). Quite how Mercury entered this '3/2 spin-orbit resonance' has been a puzzle — although now Correia and Laskar² (on page 848 of this issue) propose a solution.

The initial spin rate of our own satellite might have been as short as 10 hours, but it has been braked over time by the action of the tides raised on the Moon by Earth. Because the orbit of the Moon is eccentric, its rotational period should have ended up about 3% lower than the orbital period — with the result that, over a period of about three years, we would be permitted to see both sides of our satellite^{3,4}. But the synchronous state of matching spin and orbital rates — a 1/1 spin-orbit resonance — has been reached because the Moon has a small, permanent deformation. The gravitational interaction between the Earth and the quadrupole moment of the Moon accounts for the stability of the 1/1 spin-orbit resonance⁵.

That other spin-orbit resonances were possible was not realized before the 1965 radar observations of Mercury, made at the Arecibo Observatory in Puerto Rico. But the stability of these spin-orbit resonances was quickly explained^{6–8}. The dynamical stability of Mercury's spin state is best understood by plotting the path of the Sun in a reference frame centred on, and rotating with, the solid body of the planet (Fig. 1). Because the ratio of the rotational and orbital periods is the ratio of two integers, the path in the rotating frame is closed. Analysis shows that the oscillation of the angle between the long axis of the planet and the direction of pericentre (the point in the orbit at which the Sun is closest; Fig. 1b), follows the same equation as describes the damped oscillations of a pendulum⁴.

So how did Mercury enter this resonance? There are two terms in the equation of motion for the planet. One term describes the strength of the resonance (the depth of the potential well), which in this case depends on the eccentricity of Mercury's orbit and the resonant integers — basically, the shape of the looped path in Fig. 1b. The second term depends on the tidal torque exerted by the Sun that drives the spin towards the resonant encounter. The problem is that if these two terms remain constant, the pendulum equation is reversible

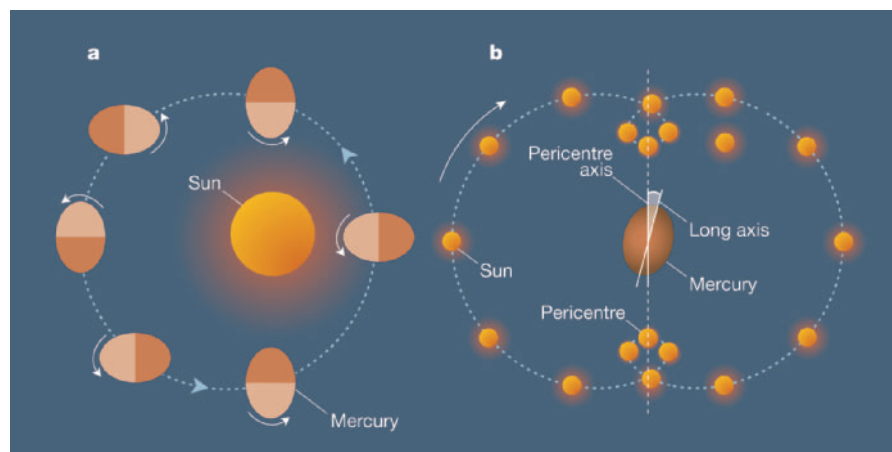


Figure 1 Mercury's 3/2 spin-orbit resonance. a, The rotational period of the planet Mercury is exactly two-thirds of its orbital period. Hence, on every second passage of the planet through the pericentre (the point in the elliptical orbit closest to the Sun), Mercury presents the same face to the Sun. b, The dynamical stability of this unusual resonant lock, or spin-orbit coupling, can be understood by plotting, at equal intervals of time, the position of the Sun in a reference frame that is centred on Mercury and rotates with the solid body of the planet. The angle described by the long axis of the planet and the direction of pericentre oscillates like a pendulum and follows the damped-pendulum equation⁴.

and the system passes through the resonance without capture⁸. Previous attempts to understand Mercury's capture into its spin-orbit resonance invoked changes in the tidal torque that broke the symmetry of the system; capture was possible, but the probability of its happening was on the low side⁸, at only 7%.

Correia and Laskar² have achieved new insight into the problem of the capture of Mercury through their investigation of the long-term dynamical evolution of the planet's orbital eccentricity (that is, how much it deviates from a perfect circle). Periodic oscillations of planetary orbital eccentricities and of their inclinations to the plane of Earth's orbit around the Sun were first analysed by Joseph Louis Lagrange, in terms of coupled linear oscillators⁹. More recent analyses of these regular oscillations suggest that the orbital eccentricity of Mercury should vary between 0.11 and 0.25 (eccentricity is zero for a circle). If that variation is included in the capture model, the probability of capture decreases — making the problem of resonant capture even worse⁴.

However, Laskar has shown in earlier work¹⁰ that the variations in the orbital eccentricities and inclinations of the inner, terrestrial planets cannot be completely described by a sum of the normal modes of coupled oscillators. In fact, the motions of these orbital elements are chaotic on timescales of millions of years; Mercury's eccentricity shows the greatest chaotic variation, from near zero to as high as 0.45 or more¹⁰. When this larger variation is factored into the capture, as Correia and Laskar have now done², it at last becomes clear how Mercury could have arrived in its 3/2 spin-orbit resonant state. Because the eccentricity can decrease to near zero, the strength of the resonant coupling can similarly drop to near zero (the looped path in Fig. 1b would be a uniform circle); all resonances except the 1/1 resonance could become unstable, allowing the planet to escape from resonance. In contrast, the excursions of the eccentricity to high values is tracked by corresponding changes in Mercury's spin rate, driven by the tidal torque. The result is that some resonant states — including the 3/2 spin-orbit resonance — are passed through many times and the probability of eventual capture is greatly increased.

Correia and Laskar's calculations² suggest that, over a four-billion-year period, the most likely state for Mercury to be captured in is, in fact, the 3/2 spin-orbit resonance. The chaotic variation of the planet's eccentricity means that this was no improbable accident after all.

Stanley F. Dermott is in the Department of Astronomy, University of Florida, Gainesville, Florida 32611-2055, USA.
e-mail: dermott@astro.ufl.edu

2. Correia, A. C. M. & Laskar, J. *Nature* **429**, 848–850 (2004).
3. Goldreich, P. *Astron. J.* **71**, 1–7 (1966).
4. Murray, C. D. & Dermott, S. F. *Solar System Dynamics* (Cambridge Univ. Press, 1999).
5. Jeffreys, H. *The Earth* (Cambridge Univ. Press, 1970).
6. Colombo, G. *Nature* **208**, 575 (1965).

7. Colombo, G. & Shapiro, I. I. *Astrophys. J.* **145**, 296–307 (1966).
8. Goldreich, P. & Peale, S. J. *Astron. J.* **71**, 425–438 (1966).
9. Brouwer, D. & Clemence, G. M. *Celestial Mechanics* (Academic, Orlando, 1961).
10. Laskar, J. *Astron. Astrophys.* **287**, L9–L12 (1994).

Condensed-matter physics

Plasmas put in order

Thomas C. Killian

Plasmas are usually a hot soup of dissociated electrons and ions. There are, however, techniques for cooling plasmas, and simulations show that an ultracold plasma could be made to crystallize.

In a typical plasma, energetic collisions tear neutral atoms apart to produce ions and electrons, which then attract or repel each other through the Coulomb force. Plasmas must be hot — as in a flame, or on the surface of the Sun — for this process to occur. At such high temperatures, the random thermal motion of the particles dominates; the positions of individual particles show no correlation or order, despite their Coulomb interactions. In *Physical Review Letters*, Pohl *et al.*¹ show, through computer simulations, how this situation might be reversed: by laser-cooling a neutral plasma, a system could be created in which the Coulomb interactions dominate and the particles arrange themselves into ordered shells or lattices.

Plasmas in which the Coulomb interaction is larger than the thermal energy are described as being strongly coupled. In nature, such plasmas are expected to exist in exotic environments, such as the crusts of neutron stars and the interiors of gas-giant planets. A few examples of strongly coupled plasmas have been created in the laboratory, such as laser-cooled ions, in Penning traps² or storage rings³, that freeze at millikelvin temperatures to form lattices called Wigner crystals. Dusty plasmas⁴ of highly charged, micrometre-size spheres suspended in a discharge plasma show similar ordering.

Laser cooling has not been used on a neutral plasma because the high energies involved would overwhelm the cooling force, or the plasma would expand into the surrounding vacuum before the lasers could do their job. Recent experiments, however, have created ultracold neutral plasmas⁵ that are cold enough to make laser cooling of the plasma feasible. To create an ultracold neutral plasma, atoms are first laser-cooled to about 1 mK and then excited by a laser pulse to an energy just above the ionization potential. The temperature of the electrons freed by ionization is roughly equal to the difference between the ionizing photon's energy and the ionization potential, and can easily be tuned from 1 to 1,000 K. The initial kinetic energy of the ions, because of their large mass, is close to that of the original neutral atoms, although

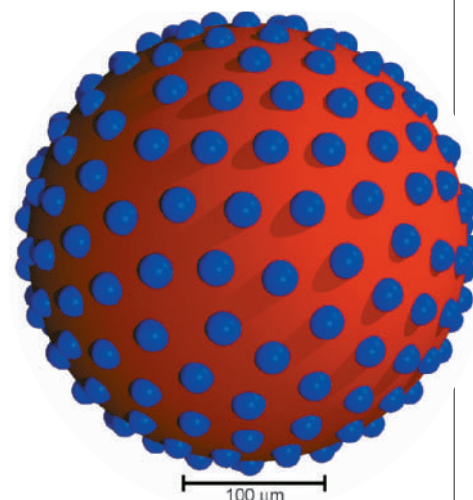


Figure 1 Crystallization in a laser-cooled neutral plasma. Simulations by Pohl *et al.*¹ indicate that the ions in a neutral plasma could take on ordered structures when subjected to laser cooling. This image shows the typical arrangement of ions in one of several concentric ion shells.

equilibration in the first few hundred nanoseconds following ionization raises the ion temperature to about 1 K. Subsequent laser cooling of the ions should push the plasma deep into the strongly coupled regime^{6,7}.

To model ultracold neutral plasmas, Pohl *et al.*¹ used a molecular-dynamics calculation to track the positions and velocities of the constituent particles as they expand into the surrounding vacuum. This would be an undergraduate physics problem if the system consisted of only one electron and one ion. But the number of interactions that must be calculated during each time-step of the simulation scales as the square of the number of particles involved. This makes it a herculean task to track the 100,000 particles necessary to capture the dynamics of an ultracold neutral plasma. Fortunately, some simplifications are possible. If the electrons are hot enough (at about 30 K), they move quickly and follow the potential-energy surface created by the positive ions. The electrons can then be treated as a background fluid whose equilibrium properties are easily calculated

1. Pettengill, G. H. & Dyce, R. B. *Nature* **206**, 1240 (1965).

at each step. This reduces the number of particles by half, but more importantly, it means that the massive, sluggish ions, not the speedy electrons, determine the minimum time-step, so a few days' calculation can cover the microsecond timescale on which ordering of the ions would occur.

Pohl and colleagues' simulations of a spherically symmetric plasma, with an initial ion density of around 10^8 cm^{-3} and a diameter of 20 μm , show that the ions first crystallize into a lattice structure as the plasma is cooled. If the cooling is rapid relative to the timescale of the plasma's expansion, the ions relax further into concentric shells, in each of which there is two-dimensional hexagonal ordering of the ions (Fig. 1).

The first step towards achieving such crystallization has already been taken, with the demonstration of a laser spectroscopic probe⁵ of ions in the plasma. Here, the temperature of the ions and the expansion of the plasma are monitored through the Doppler

shifts in the spectrum of laser radiation absorbed by the ions. Further studies along these lines will reveal the forces at work during the expansion, and the laser-ion interaction that allows spectroscopy is the same as that required for laser cooling. Pohl *et al.*¹ have developed a powerful tool with which to model crystallization in laser-cooled neutral plasmas. Now it's up to the experimenters to make it happen. ■

Thomas C. Killian is in the Department of Physics and Astronomy, Rice University, Houston, Texas 77251-1892, USA.

e-mail: killian@rice.edu

1. Pohl, T., Pattard, T. & Rost, J. M. *Phys. Rev. Lett.* **92**, 155003 (2004).
2. Itano, W. M. *et al. Science* **279**, 686–689 (1998).
3. Waki, I., Kassner, S., Birkel, G. & Walther, H. *Phys. Rev. Lett.* **68**, 2007–2010 (1992).
4. Morfill, G. E., Thomas, H. M., Konopka, U. & Zuzic, M. *Phys. Plasmas* **6**, 1769–1780 (1999).
5. Simien, C. E. *et al. Phys. Rev. Lett.* **92**, 143001 (2004).
6. Kuzmin, S. G. & O'Neil, T. M. *Phys. Plasmas* **9**, 3743–3751 (2002).
7. Killian, T. C. *et al. J. Phys. A* **36**, 6077–6085 (2003).

Cell biology

A channel for protein waste

Randy Schekman

Cells destroy misshapen proteins; viruses use the same methods to destroy healthy cellular proteins that are involved in antiviral defences. A long-sought intermediary in the process has now been uncovered.

Cells go to great lengths to ensure that protein molecules fold properly and function in the correct cellular compartment. Mistakes are dealt with harshly: the offending proteins are destroyed. On pages 834 and 841 of this issue, Lilley and Ploegh¹ and Ye *et al.*² describe how they identified a molecule that helps redirect proteins out of one compartment, the endoplasmic reticulum, to the waste-disposal machinery.

At first glance, the process of weeding out unwanted proteins seems straightforward enough. An elaborate cellular machine, the proteasome, attacks misfolded proteins that have become tagged with a small polypeptide marker, called ubiquitin. This machine is driven by the cellular energy store, ATP.

For many years, this editing function was thought to target only proteins that are found in the body of the cell—the cytoplasm—and it was assumed that other degrading enzymes would deal with proteins in distinct compartments. But around a decade ago, converging lines of investigation highlighted a role for the proteasome in the degradation of proteins that misfold in the endoplasmic reticulum (ER)^{3–5}, a major site of protein synthesis and the first port of call for proteins that are destined for the cell surface or to be secreted. Thus, mutant glycoproteins are somehow regurgitated to the cytoplasm,

where ubiquitin tagging promotes the recruitment of the proteasome to the surface of the ER. ATP then drives the ubiquitin-tagged protein into the clutches of the proteasome through the intervention of another protein, called p97 in mammals⁶. The net outcome is that damaged goods are reduced to peptides and glycans.

Certain viruses that seek to subvert the capacity of an immunologically competent cell to mount an antiviral defence have exploited this editing pathway. Proteins known as class I major histocompatibility complex (MHC) molecules are essential in alerting the immune system to the presence of viruses, but cytomegalovirus has evolved a devious means of diverting newly synthesized MHC molecules from this task. Two viral glycoproteins, US2 and US11, insert themselves into the ER membrane and interrupt the flow of class I molecules to the cell surface, redirecting them to an enzyme that is responsible for ubiquitination and thus into the jaws of the proteasome⁷.

On their way out of the ER, redirected MHC class I molecules are assumed to pass through the same portal that is used for the regurgitation of misfolded cellular proteins. But no such connection has been firmly established, nor is the identity of this portal known. One candidate for such a channel is Sec61, a protein that creates the pore through



100 YEARS AGO

It is eighteen months or more since Mr. Marconi succeeded in establishing wireless communication across the Atlantic. On that occasion a few congratulatory messages were exchanged, a great deal was written on the subject in the Press, and the more timorous of cable shareholders were reported to be much troubled. A little later the attempt was made to demonstrate that this achievement was not merely a firework display, but was capable of direct commercial application; the Marconi Co. entered into a contract to supply the *Times* with news from America by wireless telegraphy, and for a day or so there appeared items of news in that paper under the heading "By Marconigraph." But after a few messages something went wrong, and the public were given to understand that a piece of auxiliary machinery had broken down. It is to be presumed that this piece of machinery has at length been repaired, for Mr. Marconi has once again come very much to the front with long-distance transmission work. The announcement, which we published last week, that he had been successful in maintaining a supply of news to the *Campania* on her voyage across the Atlantic with a regularity sufficient to allow of the publication of a daily paper on board that vessel affords evidence that he is still steadily pushing forward the practical development of wireless telegraphy. We have repeatedly urged in these columns that the real work of wireless telegraphy lay in communication with ships, and it is therefore a greater pleasure to record this latest development than it would be to announce the reopening of Transatlantic communication.

From *Nature* 23 June 1904.

50 YEARS AGO

In the House of Commons on June 15, Mr. Geoffrey de Freitas asked the Under-Secretary for Air whether the physical sub-committee of the Meteorological Research Committee had yet considered the problem of weather modification; and what conclusions it reached... Mr. George Ward, in a written reply, stated that the committee had recently considered this subject and come to the conclusion that there is no reliable evidence that rainfall has ever been artificially increased on an economically useful scale, and that there is no scientific basis for believing that any method yet proposed would be successful in achieving such a result.

From *Nature* 26 June 1954.

which secretory and membrane proteins become inserted into the ER during their biosynthesis. Numerous experiments have hinted that Sec61 molecules transiently interact with MHC class I molecules as they are diverted to the proteasome⁸, and that mutations in Sec61 retard the degradation of misfolded secretory proteins^{9,10}. Unfortunately, these experiments did not reveal a direct molecular contact between Sec61 and proteins being redirected out of the ER. Other genetic studies have identified other membrane proteins that participate in the degradation of misfolded secretory proteins^{11,12}. But, until now, no evidence linked these molecules directly to the transport of proteins out of the ER.

The two reports in this issue^{1,2} highlight the role of a mammalian equivalent of the yeast Der1 protein — which is found in the ER membrane and is required for the degradation of certain misfolded glycoproteins¹¹ — in removing proteins from the ER. Lilley and Ploegh¹ used the cytomegalovirus US11 protein to probe the environment of MHC class I proteins as they are diverted from the ER. By using tagged forms of wild-type and non-functional US11, the authors could track molecules in the ER membrane. They found that several previously unknown proteins could be precipitated in a complex with wild-type, but not mutant, US11.

One subunit of this complex was one of several mammalian relatives of yeast Der1. Like its yeast counterpart, this mammalian protein, Derlin-1, is found in the ER membrane, with the protein's chain snaking back and forth across the membrane a possible four times. It binds to wild-type US11 and MHC, but the complex decomposes quickly unless the proteasome is chemically inhibited.

A crucial experiment, documenting the physiological importance of Derlin-1, came with the demonstration that a hybrid protein containing Derlin-1 and green fluorescent protein blocks the US11-dependent degradation of MHC molecules. Surprisingly, though, this hybrid does not interfere with the action of US2, the other cytomegalovirus protein that diverts MHC molecules from the ER. So, US11 and US2 have distinct mechanisms of action. An obvious possibility is that US2 acts through one of the other mammalian Derlin proteins. In uninfected cells, each Derlin might proofread a restricted cohort of molecules.

Meanwhile, Ye *et al.*² pursued a different path to discover the ER proteins involved in MHC degradation. Given an essential role for the ATP-driven molecule p97, the authors probed the ER membrane for a p97 receptor. Their efforts were rewarded with a complex of two proteins, one of which was Derlin-1 and one of which they named VIMP (which might be one of the other unknown proteins that Lilley and Ploegh found).

Derlin-1, VIMP and p97 co-localize to the ER, and might constitute the complex to which US11 and MHC class I molecules are recruited to initiate the export event. Using a cell-free reaction that reproduces MHC export¹³, Ye *et al.* trapped an intermediate consisting of VIMP, Derlin-1, US11, ubiquitinated MHC class I protein and p97. The stable formation of this complex required functional US11.

Broadening their findings, Ye *et al.* also found that when cells were treated with dithiothreitol, a reducing agent that causes misfolded proteins to be directed out of the ER, more newly synthesized proteins that were bound up with VIMP accumulated. And blocking Derlin-1 expression in nematode worms promoted a form of cellular stress referred to as the unfolded-protein response, consistent with a general problem in the disposal of misfolded proteins.

So these studies^{1,2} highlight a role for the Derlin-1 protein, and possibly for other members of the gene family, in the selective extrusion of undesirable proteins from the ER. It is probably not the only such channel, however; yeast with mutations in Der1 can still dispose of many mutant membrane

proteins, for instance^{14,15} (R. Hampton, personal communication). Cells clearly hold many more secrets about how they deal with malfunctioning parts.

Randy Schekman is in the Department of Molecular and Cell Biology, Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA.

e-mail: schekman@berkeley.edu

1. Lilley, B. N. & Ploegh, H. L. *Nature* **429**, 834–840 (2004).
2. Ye, Y., Shibata, Y., Yun, C., Ron, D. & Rapoport, T. A. *Nature* **429**, 841–847 (2004).
3. Sommer, T. & Jentsch, S. *Nature* **365**, 176–179 (1993).
4. Hiller, M. M., Finger, A., Schweiger, M. & Wolf, D. H. *Science* **273**, 1725–1728 (1996).
5. McCracken, A. A. & Brodsky, J. *J. Cell Biol.* **132**, 291–298 (1996).
6. Ye, Y., Meyer, H. H. & Rapoport, T. A. *Nature* **414**, 652–656 (2001).
7. Weitz, E. J. *et al.* *Cell* **84**, 769–779 (1996).
8. Wiertz, E. J. *et al.* *Nature* **384**, 432–438 (1996).
9. Pilon, M., Schekman, R. & Römisch, K. *EMBO J.* **16**, 4540–4548 (1997).
10. Plemper, R. K., Bohmler, S., Boddallo, J., Sommer, T. & Wolf, D. H. *Nature* **388**, 891–895 (1997).
11. Knop, M., Finger, A., Braun, T., Hellmuth, K. & Wolf, D. H. *EMBO J.* **15**, 753–763 (1996).
12. Hampton, R., Gardner, R. G. & Rine, J. *Mol. Biol. Cell* **7**, 2029–2044 (1996).
13. Shamu, C. E., Story, C. M., Rapoport, T. A. & Ploegh, H. L. *J. Cell Biol.* **147**, 45–58 (1999).
14. Hill, K. & Cooper, A. A. *EMBO J.* **19**, 550–561 (2000).
15. Vashist, S. & Ng, D. T. *J. Cell Biol.* **165**, 41–52 (2004).

Cometary science

Fly-through at Wild 2

Michael F. A'Hearn

The Stardust mission has made the closest approach ever to a comet. Its 'fly-through' of the gas and dust surrounding Wild 2 presents a unique opportunity to investigate the evolution of such bodies.

On 2 January this year, the Stardust spacecraft flew through the coma of the comet Wild 2. The primary purpose of this NASA mission is to collect particles from the coma — the dust and gas escaping from the comet's solid nucleus — and return them to Earth. The sample return is scheduled for January 2006, but in the meantime the dramatic results from the fly-through itself are now reported in *Science*^{1–4}.

Two other comets have been visited by spacecraft — Halley in 1986 and Borrelly in 2001. But all the measurements made at Wild 2 seem very different from those taken at the other two comets. Apart from the data on dust flux, it seems that these differences cannot be explained by the fact that the distance of closest approach of Stardust to Wild 2 was much smaller — only 236 km — than in the previous cometary fly-bys. But there is a major difference between Wild 2 and the other two comets: Halley and Borrelly have been in their present orbits, with only small changes, for about 100 orbital periods, which is as far back as reliable orbital calculations can be made; Wild 2 has not.

In 1974, Wild 2 passed very close to

Jupiter and was kicked by the planet's gravity into an orbit much closer to the Sun. Its orbital period dropped from 40 years to about 6 years. Today, Wild 2's perihelion (its point of closest approach to the Sun) is roughly equivalent to the size of Earth's orbit (that is, one astronomical unit, or about 150 million kilometres). At such a distance, heat from the Sun is likely to cause changes in the comet's surface and composition. Although calculations of the comet's past orbit are not accurate enough to tell us whether it ever previously had such a small perihelion distance, most astronomers assume that Wild 2 has had only a few perihelion passages close enough to cause significant evolution of the comet — unlike Halley and Borrelly. Wild 2 should be much less evolved than either of these comets.

So what are the differences seen in these early results from Stardust? In the first of the new papers, Brownlee *et al.*¹ report that the 5-km nucleus of Wild 2 is much more nearly spherical than are the nuclei of either of the other comets (Fig. 1). This would tend to support the idea that 'outgassing', which peaks at perihelion and is very sensitive to the

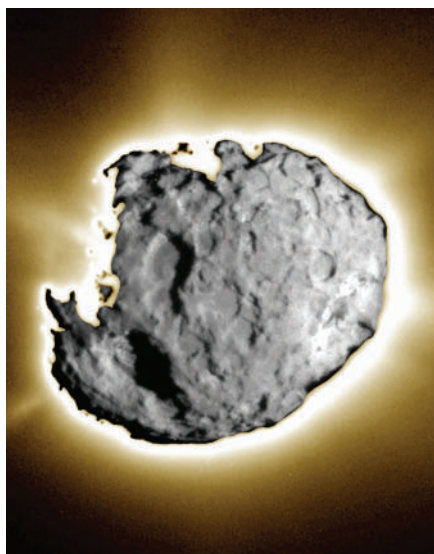


Figure 1 Close-up on Wild 2. This composite image from the Stardust spacecraft shows the cratered surface of the comet and its halo of gas and dust, which seems collimated into several jets.

distance from the Sun, might tend to erode the 'waist' of a nucleus, making more evolved comets more prolate.

The topographic relief on Wild 2, however, is substantially greater (at least 200 m)¹ than that on Borrelly (about 100 m), even though both comets have similar volumes and thus similar gravitational states (assuming that their densities are comparable). The slopes on Wild 2 are steep — nearly vertical in some cases. These facts, and the presence of at least one prominent overhang, suggest that the surface layers are strong and cohesive, and might be expected to get stronger with additional evolution. And yet the features on the presumably more evolved comet Borrelly are not suggestive of strength in its surface material. Borrelly has large mesas (raised plateaux), which have been interpreted as being eaten away by outgassing along the mesa walls⁵. Wild 2 has a much smaller area of mesas — which would, contrarily, suggest that Wild 2 is the more evolved of the two, with most of its mesas having been eaten away. But then the existence of preserved craters on Wild 2 and not on Borrelly implies that the opposite is true.

There are many features — and more mysteries — on the surface of Wild 2, including circular features that Brownlee *et al.*¹ attribute to impacts, and two footprint-like depressions that the Stardust team have named Left Foot and Right Foot. However, there are no small features (less than 0.5 km in diameter) that could be associated with impacts, and there seems to be no displaced material left from larger impacts (although it may be that the escape velocity on Wild 2 is sufficiently low that this material has left the surface). If the circular features are indeed impact craters, the surface must, overall, be rather old for there to have been so many

impacts. Other non-circular features seem to be associated with streams of gas, or jets, in the coma (Fig. 1), produced by a combination of outgassing and other processes and modelled in the second of the papers, by Sekanina *et al.*². For these sublimation vents not to have completely obliterated the comet surface, the jets must have developed fairly recently — perhaps they have been active only since the decrease in perihelion distance 30 years ago.

The dust counters on board Stardust, monitored by Tuzzolino *et al.*³, picked up narrow bursts of dust particles during the fly-through, with very sharp transitions between a high density of particles and essentially no particles. This may be caused by the spontaneous fragmentation of dust grains in the inner coma. The fragments would have dispersed by the time they reach the outer coma, which would explain why the Vega spacecraft that flew past comet Halley at a greater distance did not detect the same effect. The sharp transitions might also have been disguised by the more limited resolution of Vega's instruments and its higher fly-by speed. Using stereoscopy in optical images of the jets, Sekanina *et al.*² have determined their three-dimensional orientations and successfully correlated them with some of the individual bursts in the dust.

Finally, the chemical analysis of the dust particles presented by Kissel *et al.*⁴ has

revealed marked differences compared with a similar analysis for comet Halley. Most notable is the almost complete absence of atomic ions in the mass spectra at Wild 2, perhaps due in part to the much lower fly-by speed (6 km s⁻¹ compared with 68 km s⁻¹, relative to each comet) and thus smaller impact speed of the dust particles onto Stardust's monitors. As at Halley, organic molecules are abundant, but the material at Wild 2 seems more nitrogen-rich. Generally, the grains at Wild 2 seem depleted in C=O-bonded material compared with the grains at Halley. Kissel *et al.*⁴ also exclude the presence of water ice and free amino acids.

Although these papers¹⁻⁴ represent only a preliminary analysis of the data from Wild 2, the results are already intriguing. I look forward to the more detailed analyses to come — which will help to identify the evolutionary and intrinsic differences between the comets — and, of course, to Stardust's return to Earth in 2006.

Michael F. A'Hearn is in the Department of Astronomy, University of Maryland, College Park, Maryland 20742-2421, USA.
e-mail: ma@astro.umd.edu

1. Brownlee, D. E. *et al. Science* **304**, 1764–1769 (2004).
2. Sekanina, Z., Brownlee, D. E., Economou, T. E., Tuzzolino, A. J. & Green, S. F. *Science* **304**, 1769–1774 (2004).
3. Tuzzolino, A. J. *et al. Science* **304**, 1776–1780 (2004).
4. Kissel, J., Krueger, F. R., Silén, J. & Clark, B. C. *Science* **304**, 1774–1776 (2004).
5. Britt, D. T. *et al. Icarus* **167**, 45–53 (2004).

Heart disease

An ongoing genetic battle?

Deepak Srivastava

Babies born with physical defects in their hearts may survive, but they often suffer defects in heart function as adults. The physical and functional problems might, it seems, have the same genetic cause.

Thirty years ago, children born with a heart defect had little chance of survival. Thanks to surgical and medical advances, the outlook now is usually not so grim, such that although the incidence of congenital heart disease (CHD) has not changed substantially (still nearly 1 out of every 100 live births), its prevalence is rising rapidly¹. In fact, there are now more than 1 million survivors of CHD in the United States alone. But as these people enter their teens and twenties, new disease processes are becoming apparent, including abnormal conduction of electricity within the heart and a diminished ability of the heart muscle to contract. These 'secondary' defects have generally been ascribed to abnormal blood flow resulting from the primary defect, but Pashmforoush *et al.*², writing in *Cell*, suggest that the same mutated gene that causes an early defect in heart development in mice might later be directly involved in cardiac

dysfunction, even in adulthood. If true in humans, the finding might allow postnatal approaches to alter the natural history of CHD in genetically at-risk individuals.

Pashmforoush and colleagues² studied a master regulatory protein called Nkx2-5, which controls many other genes and is essential for normal heart formation in fruit-flies, mice and humans³⁻⁵. People with mutations in one of their two copies of this gene typically show abnormal communication between the two atrial chambers of the heart (an atrial septal defect), and progressive disruption of electrical conduction through the cardiac chambers, which can result in sudden death⁵. Although the mechanisms remain unclear, this protein is known to interact with two others, Tbx5 and Gata4, mutations in which cause a similar CHD in humans⁶⁻⁸.

Previous work on the Nkx2-5 protein elegantly illustrated the usefulness of cross-species studies to inform biologists

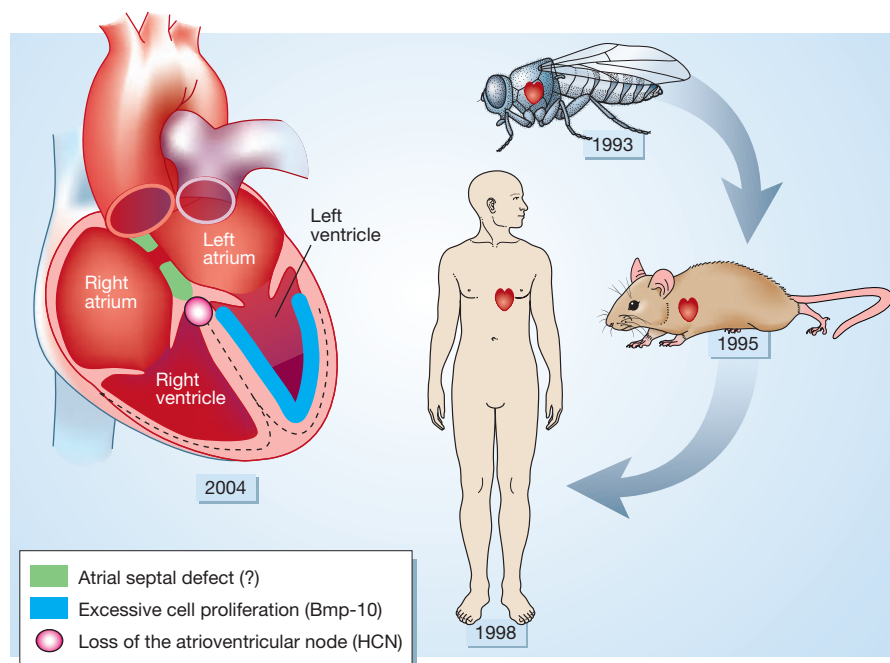


Figure 1 Problems of the heart in young and old. Clockwise from top, the Nkx2-5 protein was discovered to be required for normal heart development first in fruitflies³, then mice⁴ and finally humans⁵. For instance, mutations in Nkx2-5 lead, by an unknown mechanism, to a defect in the wall between the two atrial chambers of the heart (an atrial septal defect; left). Pashmforoush *et al.*² have now found that Nkx2-5 mutations also produce defects in the hearts of adult mice. For instance, such mutations (perhaps by increasing the concentrations of the Bmp-10 protein) lead to increased cell proliferation in the wall of the ventricles. Possibly by increasing the concentrations of the channel protein HCN, they also lead to progressive loss of the atrioventricular node. Finally, deregulation of sarcolipin might lead to decreased cardiac function (not shown).

and clinicians about disease characteristics and mechanisms, and the new paper continues that theme. Using sophisticated genetic technologies, Pashmforoush *et al.* engineered mice to lack Nkx2-5 specifically in muscle cells in the ventricular chambers of the heart, after the initial muscle specification event had occurred.

They found that the animals did not die, as they would if they lacked Nkx2-5 in all tissues. Surprisingly, however, they suffered as adults from abnormally thickened ventricles and disruption of the conduction system. Unlike in the humans described above, this mutation did not cause gross morphological defects in the heart at birth, except that a conduction centre called the atrioventricular node — the key site of electrical communication between the atria and ventricles — was smaller than normal. But although this node allowed conduction at first, the specialized muscle-derived cells in the node were lost over time and replaced by scar tissue, resulting in progressive defects in electrical conduction.

Because Nkx2-5 was disrupted solely in the ventricular muscle lineage, this work clearly demonstrates that the formation and maintenance of the atrioventricular node depend on Nkx2-5 functioning in heart muscle cells, rather than in other heart cell types. In addition, the investigators show that this protein is involved in postnatal

regulation of the function and proliferation of cardiac muscle cells — all of which, they find, were similarly affected in a patient with a mutation in this gene.

The insights of this work come not from the ability to copy a human disease in mice, but rather through the use of a mouse model of that disease to understand the underlying mechanisms. Using state-of-the-art bioinformatics and microarray technologies, Pashmforoush *et al.* found, and confirmed the identity of, numerous genes that are deregulated in hearts lacking Nkx2-5. Their data suggest that this protein normally inhibits the production of another, bone morphogenetic protein-10 (Bmp-10), and that, in the absence of Nkx2-5, concentrations of Bmp-10 increase — perhaps causing the thick ventricles seen in mice and humans with Nkx2-5 mutations (Fig. 1). Similarly, the authors propose that a deregulation of contractile genes in the muscle and ion channels in conduction tissue contributes to the progressive dysfunction of muscle and the atrioventricular node, respectively. This work therefore provides a framework for beginning to understand how cardiac developmental genes might be co-opted for additional tasks, extending into normal heart maintenance and function throughout life.

So this work represents a milestone in our understanding of the long-term consequences of CHD, but of course questions

remain. For example, the authors assume that the progressive loss of conduction cells in the absence of Nkx2-5 means that Nkx2-5 has a role in maintaining those cells. But it is also possible that there is ongoing turnover of muscle-derived conduction cells, and that, in the absence of Nkx2-5, a putative pool of progenitor cells is unable to replenish the ageing conduction cells — resulting in progressive abnormalities. In addition, although the authors propose that inhibiting Bmp-10 might be useful in preventing the thick ventricles observed when Nkx2-5 is disrupted, the current evidence remains circumstantial and awaits genetic or chemical proof of a direct role for Bmp-10 in mediating the effects of Nkx2-5 mutations.

Finally, the absence of atrial septal defects in Pashmforoush and colleagues' mutant mice could be because Nkx2-5 has a role in an alternative lineage, such as non-muscle cells derived from the endocardial lining of the heart, or it could simply reflect the fact that the 80–90% decrease in Nkx2-5 concentrations was above the threshold to produce such a morphogenetic defect. Nonetheless, the findings show that the functional anomalies seen in patients with Nkx2-5 mutations are not secondary to the morphological defects.

The discovery that one gene has distinct roles in developing and adult hearts has implications for our understanding of the natural history of CHD. Patients who survive CHD as a result of surgical intervention often have good long-term outcomes, but others suffer complications ranging from ventricular dysfunction to sudden death. Studies so far have failed to identify clinical or other predictors of outcome in such patients, and the events are thought to be stochastic.

Nevertheless, with the increasing recognition that CHD has a significant genetic component⁹, it becomes plausible to imagine that genetic variations underlie both the initial morphogenetic defects and the predisposition to long-term consequences. For instance, many adult-onset cardiac rhythm disturbances, previously considered to be 'acquired' throughout life, might in fact be the result of a genetic mutation that renders portions of the conduction system 'programmed' to fail over time, and so might be congenital in nature. If so, efforts to identify the genetic variations will be essential, as therapeutic or preventive measures, throughout childhood and in adulthood, might then be possible. It is probably time to reconsider the idea of the heart as a static organ and to recognize that previous distinctions between childhood and adult-onset heart disease might in fact be hindering us from getting to the heart of the problem. ■

Deepak Srivastava is in the Departments of Pediatrics (Cardiology) and Molecular Biology, University of Texas Southwestern Medical Center and Children's Medical Center Dallas,

6000 Harry Hines Boulevard, Dallas,
Texas 75390-9148, USA.
e-mail: deepak.srivastava@utsouthwestern.edu

1. Hoffman, J. I. E. & Kaplan, S. *J. Am. Coll. Cardiol.* **39**, 1890–1900 (2002).
2. Pashmforoush, M. *et al. Cell* **117**, 373–386 (2004).

3. Bodmer, R. *Development* **118**, 719–729 (1993).
4. Lyons, I. *et al. Genes Dev.* **9**, 1654–1666 (1995).
5. Schott, J. J. *et al. Science* **281**, 108–111 (1998).
6. Basson, C. T. *et al. Nature Genet.* **15**, 30–35 (1997).
7. Bruneau, B. G. *et al. Cell* **106**, 709–721 (2001).
8. Garg, V. *et al. Nature* **424**, 443–447 (2003).
9. Srivastava, D. *Annu. Rev. Physiol.* **63**, 451–470 (2001).

Metrology

Electrifying effects in colloids

Patrick Warren

The electric field generated in the sedimentation or centrifugation of charged colloidal particles could be exploited to determine the charge and the mass of macromolecules in a single experiment.

It might be thought that the last word on sedimentation or centrifugation had been said long ago. But over the past decade there have been persistent indications from experiment¹, theory² and simulations³ of an unusual phenomenon in the sedimentation equilibrium of suspensions of charged colloidal particles when the ionic strength is low. Colloids are small particles, typically less than one micrometre in size, that have many technological applications as well as relevance to fundamental science. On page 857 of this issue, Raşa and Philipse⁴ present convincing evidence that a macroscopic electric field, generated when the charged colloid is centrifuged to equilibrium, is behind the strange effect.

In 1926, Jean Baptiste Perrin was awarded the Nobel Prize in Physics “for his work on the discontinuous structure of matter, and especially for his discovery of sedimentation equilibrium”⁵. What Perrin had discovered was that the same law that governs the

rarefaction of Earth’s atmosphere with height — its barometric profile — also governs the distribution of suspended colloid particles undergoing brownian motion. At the time, Perrin’s measurements of the barometric profile in a painstakingly prepared gamboge (gum resin) suspension gave an independent determination of Avogadro’s number, thus contributing to the establishment of the atomistic hypothesis as experimental fact. In centrifugation — which is now well established as a means of determining molecular mass — the modest effects of Earth’s gravity are replaced by a strong radial centripetal acceleration, as a sample is whirled around at high speed in a specialist device.

Piazza *et al.*¹ found that, for charged colloidal particles at reduced ionic strength, the barometric profile is inflated, as though the particles weigh less than they should do. Biben and Hansen² have since suggested, using density functional theory, that the sample column behaves like a condenser, in

the sense that imbalanced charges accumulate at the top and bottom. The resulting electric field spans the whole column and acts to lift the colloid particles up against the force of gravity.

So, what is going on? A simple explanation, first discovered by van Roij⁶, relates the phenomenon to a now-classic piece of physical chemistry. In 1911, Frederick Donnan considered the equilibrium across a semi-permeable membrane separating a salt solution from a suspension of charged colloidal particles or macromolecules⁷. The membrane is permeable to the small ions of the salt but impermeable to the colloids. Naively, one might expect that the small ions would distribute themselves so as to have the same concentration on both sides of the membrane. However, Donnan discovered that this is not the case. Rather, the ions become redistributed; for example, ions with the same charge as the colloids tend to be expelled from the colloid-containing compartment — the Donnan common-ion effect (Fig. 1). The effects arise from a microscopic charge imbalance that builds up in the vicinity of the membrane, creating a potential difference between the compartments, now known as the Donnan potential.

Raşa and Philipse⁴ have made a connection between the Donnan membrane problem and the sedimentation profile of charged colloids. The macroscopic electric field in the sedimentation case can be attributed to a gradient in the Donnan potential, corresponding to the gradient in the concentration of colloidal particles (Fig. 1). The full problem has to be solved self-consistently, to obtain the coupled density and electric-field profiles — which is exactly what Raşa and Philipse have done.

Their calculations show that a careful analysis of the density profile for a dilute suspension of charged colloids or macromolecules could allow both the molecular mass and the charge of the sedimenting species to be determined. What remains to be demonstrated is that the phenomenon can be measured not just for colloidal particles but also for macromolecules. In this respect, globular proteins such as lysozyme might be good candidates. If it works, the approach offers an interesting alternative to electrophoretic methods for characterizing proteins.

Patrick Warren is at Unilever R&D Port Sunlight, Bebington, Wirral CH63 3JW, UK.

e-mail: patrick.warren@unilever.com

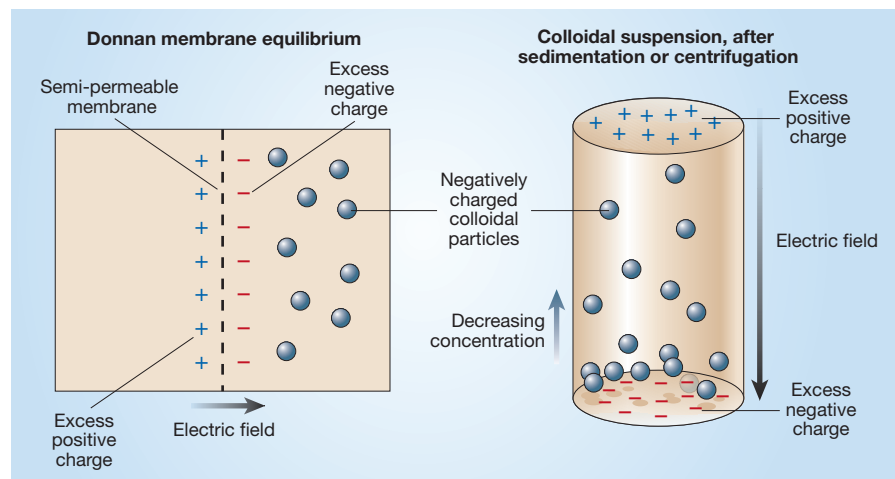


Figure 1 The Donnan membrane equilibrium and sedimentation of a charged colloidal suspension. In a Donnan equilibrium, the charge imbalance in the vicinity of a semi-permeable membrane gives rise to an electric field, with a jump in the electrostatic potential typically occurring over a length of less than a micrometre⁸. Similarly, if a colloidal suspension has a gradient of concentration (such as is produced in sedimentation or centrifugation), then a macroscopic electric field is generated by the charge imbalance appearing at the top and bottom of the sample column. The presence of this field has observable consequences for the density profile of negatively charged colloidal particles, as has been confirmed experimentally by Raşa and Philipse⁴.

1. Piazza, R., Bellini, T. & Degiorgio, V. *Phys. Rev. Lett.* **71**, 4267–4270 (1993).
2. Biben, T. & Hansen, J.-P. *J. Phys. Condens. Matter* **6**, A345–A349 (1994).
3. Hynninen, A.-P., van Roij, R. & Dijkstra, M. *Europhys. Lett.* **65**, 719–725 (2004).
4. Raşa, M. & Philipse, A. P. *Nature* **429**, 857–860 (2004).
5. *Nobel Lectures on Physics 1922–1941* (World Scientific, Singapore, 1998).
6. van Roij, R. *J. Phys. Condens. Matter* **15**, S3569–S3580 (2003).
7. Donnan, F. G. Z. *Elektrochem.* **17**, 572–581 (1911).
8. Sear, R. P. & Warren, P. B. *J. Chem. Phys.* **117**, 8074–8079 (2002).

Geochemistry

Briny reservoir salts
Vostok iceEarth Planet. Sci. Lett. **222**, 751–765 (2004)

Lake Vostok lies nearly 4,000 m beneath the Antarctic ice cap, and is thought to be the largest subglacial lake in the world. Ice samples have been taken down to a depth about 100 m above the water surface, dating back more than 400,000 years. The upper part of the ice cap is a moving glacier, but 'accretion ice', formed by the refreezing of lake water, is found below 3,538 m.

M. De Angelis *et al.* now report the first comprehensive study of the chemistry of this accretion ice. Below 3,609 m, in the layer closest to the lake, the ice contains very little salt indeed, confirming the widely predicted freshwater status of Vostok. But, surprisingly, De Angelis *et al.* find that accretion ice in the layer above (3,538–3,608 m deep) is up to 50 times saltier than normal glacier ice; it is loaded with sodium chloride, calcium sulphate and magnesium sulphate.

The authors believe that the minerals are periodically drawn into the ice from a reservoir that is linked to a shallow bay upstream from Lake Vostok. The bay is separated from the main lake by a rock outcrop, and its deeper sediments may have originally accumulated from the evaporation of ancient sea water, millions of years before Lake Vostok disappeared beneath the ice.

Mark Peplow

Parkinson's disease

Triggers of neuron suicide

Development **131**, 3229–3236 (2004)

In Parkinson's disease, neurons producing the neurotransmitter dopamine wither away in a brain area called the substantia nigra. Lavinia Albéri *et al.* present evidence that two developmental genes, *Engrailed1* and *Engrailed2*, may be involved in this process by triggering cell suicide in mice.

Both genes are active in the dopamine-producing neurons soon after the cells are made and throughout adulthood. In mice genetically engineered to lack them, however, these neurons are generated in the growing brain but do not survive beyond the birth of the animal.

Albéri *et al.* show that, in the mutant mice, the dopamine-producing neurons have committed cell suicide by roughly two-thirds of the way through embryo development. Neurons grown *in vitro* die within 24 hours of both *Engrailed* genes being switched off.

The death of these neurons in the mouse embryos occurs far earlier than their decline during Parkinson's disease in humans, which typically starts at

age 50 or older. But the authors speculate that more subtle changes in the expression levels of *Engrailed* genes might be involved in triggering neuronal death during the disease.

Helen Pearson

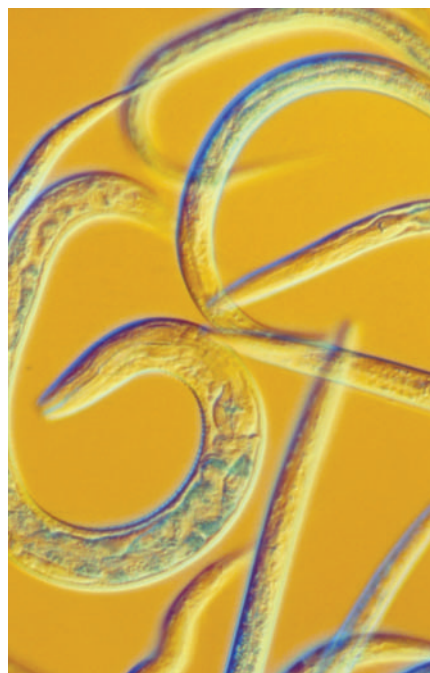
Physiology

Worms on the fumes

Neuron **42**, 731–743 (2004)

As part of the search for any genetic predisposition to alcoholism, Andrew G. Davies and colleagues work with an unlikely ally — the nematode worm *Caenorhabditis elegans*. They now report the identification of natural variation in a gene that affects the nematodes' tolerance to alcohol.

Davies *et al.* studied two strains of worm — N2 (from England) and CB4856 (from



Hawaii) — and exposed them to alcohol fumes. The resulting concentrations in the worms' bodies were the same as those that cause intoxication in humans. The Hawaiian worms recovered more quickly from the imposed binge: after 50 minutes they moved 40% as fast as sober worms, whereas the English worms managed a mere 20% of normal speed.

Genetic analysis revealed that the English worms had higher activity levels of a receptor protein called NPR-1. Moreover, mutant English worms that lacked only this gene matched their Hawaiian counterparts in their speed of recovery.

The human equivalent of NPR-1 is the receptor for neuropeptide Y. This neuropeptide is implicated in regulating a wide variety of behaviours, and the work of Davies *et al.* adds to evidence that it is somehow involved in the development of acute tolerance to alcohol, a risk factor for alcoholism.

Laura Nelson

Cancer

TAL order

Cancer Cell **5**, 587–596 (2004)

T-cell acute lymphoblastic leukaemia (T-ALL) is a notoriously treatment-resistant type of blood-cell cancer. More than 60% of T-ALL patients express a gene called *TAL1/SCL*, which is thought to trigger their condition, although the mechanism involved has been unclear. Jennifer O'Neil *et al.* now report that the gene's protein product causes leukaemia by indirectly disrupting immune-cell development.

Mice genetically engineered to produce *tal1/scl* protein show impaired immune-cell development through an inability to express certain genes, and more than 80% of the animals go on to develop the disease. *TAL1/SCL* represses genes that control immune-cell development by interfering with the action of a transcription factor called E47/HEB. It does this through the action of a protein complex, the histone deacetylase repressor complex (HDAC). Drugs that block HDAC slow the growth of isolated tumour cells, causing many to die.

The authors speculate that T-ALL patients who express *TAL1/SCL* may also be responsive to HDAC inhibitors, so the finding provides cause for further investigations.

Helen R. Pilcher

Photovoltaics

Harnessing leaf power

Nano Lett. **4**, 1079–1083 (2004)

Could your laptop run on spinach? Perhaps. Rupa Das *et al.* have isolated photosystem I, a central engine of photosynthesis, from chloroplasts of spinach leaves and tethered it to a thin film of gold deposited on a transparent, electrically conductive material. They find that the photosystem generates an electric current in response to visible light, thereby acting as a photovoltaic cell.

The photosystem is a huge, many-molecule assembly, comprising 12 core protein subunits and hundreds of chlorophyll molecules. But it apparently remains functional when deposited on the substrate via a polyhistidine linker in the presence of peptide surfactants (which presumably surround the membrane-protein assembly to provide essential stabilization). The same trick works for the bacterial reaction centre from the photosynthetic purple bacterium *Rhodospirillum rubrum*, a much simpler molecular assembly.

Das *et al.* hope ultimately to achieve photovoltaic power-conversion efficiencies of around 20%, which would be comparable to the best commercial inorganic solar cells.

Philip Ball

Hibernation in a tropical primate

Even in the wound-down hibernating state, this lemur can warm up without waking up.

The Madagascan fat-tailed dwarf lemur, *Cheirogaleus medius*, hibernates in tree holes for seven months of the year, even though winter temperatures rise to over 30 °C. Here we show that this tropical primate relies on a flexible thermal response that depends on the properties of its tree hole: if the hole is poorly insulated, body temperature fluctuates widely, passively following the ambient temperature; if well insulated, body temperature stays fairly constant and the animal undergoes regular spells of arousal. Our findings indicate that arousals are determined by maximum body temperatures and that hypometabolism in hibernating animals is not necessarily coupled to a low body temperature.

Temperate and Arctic hibernators in deep burrows are buffered against fluctuations in cold winter temperatures¹. Tropical animals, on the other hand, may use poorly insulated sites such as tree holes² and so face the problems of recurrent high temperatures and wide daily fluctuations in temperature during the tropical winter.

Our field study of *C. medius* reveals that its body temperature (T_b) during hibernation varies to an extent previously unknown in mammals (for methods, see supplementary information). Most lemurs showed a wide daily fluctuation in T_b of almost 20 °C, which closely followed the air temperature of their tree holes (T_h) in range and rate of change (Fig. 1a); the greatest fluctuation was 24.9 °C, the lowest recorded T_b was 9.3 °C and the highest was 35.9 °C.

The daily ranges and heating rates of T_b and T_h were not significantly different (Table 1), and the difference between T_b and T_h was usually very small (1.81 ± 0.91 °C; $N = 15$, $n = 16,560$ where N was the number of animals tested and n the number of data points). This passive thermal response to T_h continued over many weeks or even months and the lemurs remained ectothermic as long as T_b regularly exceeded 30 °C. At no point was hibernation interrupted by periodic euthermic arousals. Such arousals are energetically very expensive, last for 12–24 hours^{3,4}, and were previously considered

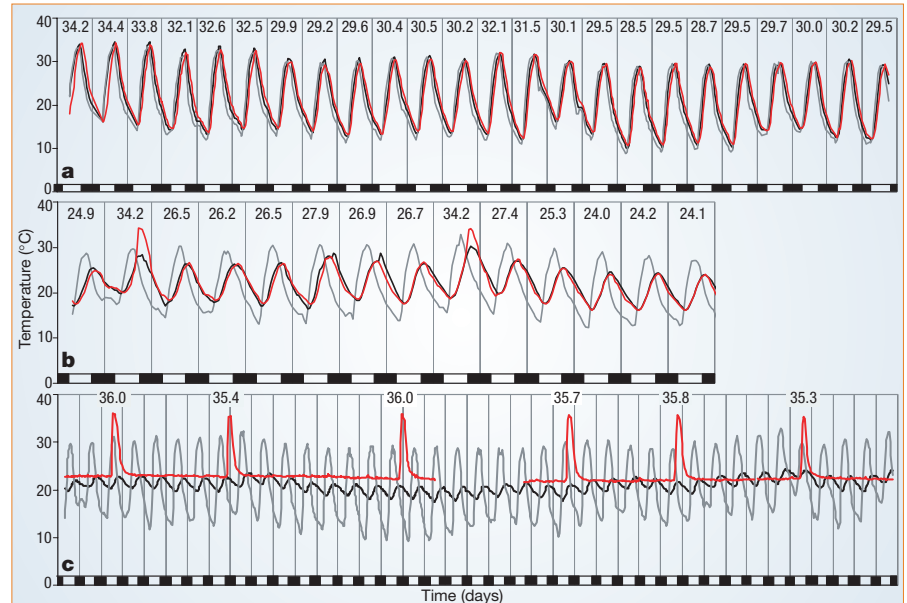


Figure 1 Body temperature of *Cheirogaleus medius* during the hibernation period. The animal's temperature (red traces) was monitored continuously in tree holes that were **a**, poorly insulated, measured over 24 d; **b**, moderately well insulated, measured over 14 d; or **c**, well insulated (*Commiphora guillaumini*; 18 d and 17 d; tree-hole temperature was measured nearer the edge of the hole and so shows greater fluctuations than animal temperature). Vertical lines, midnight; black bars, dark phase. Black traces, tree-hole temperature; grey, ambient temperature. Numbers (top) give the daily maximum body temperature (**a**, **b**) or the maximum body temperature during arousals (**c**).

to be obligatory during hibernation^{5,6}.

Some lemurs hibernated in tree holes that were better insulated; their T_b stayed below 30 °C because the ambient temperature fluctuated less. Occasionally, these animals actively raised their T_b above 30 °C after the T_h had increased T_b to its daily maximum (Fig. 1b).

Other lemurs hibernated in well insulated holes in large, thick-walled (over 20 cm) trees (*Commiphora guillaumini*) where there were only minor fluctuations in T_h during the day and T_b stayed at about 25 °C for many days (Fig. 1c). Unlike the other lemurs, these animals had an arousal with an increase in T_b about once a week (6.7 ± 3.9 d). Compared with arousals in temperate and Arctic hibernators, however, these were short: T_b was maintained above 33 °C for less than 6 h.

The amplitude of T_b during arousals was comparable to the daily T_b amplitude of passively thermoregulating lemurs in poorly

insulated tree holes (Table 1). However, the spontaneous arousals must have been induced by endogenous heat production because the rate of warming was more than double that caused by T_h . Endogenous heating becomes unnecessary when ectothermic fluctuations in T_b reach 30 °C. This indicates that arousal is determined by body temperature; 30 °C must be exceeded, at least episodically, for physiological homeostasis.

These alternative patterns of thermoregulation are not specific to individual lemurs. Some changed their hibernaculum from well to poorly insulated tree holes within the same hibernation season. They could switch from a constant T_b with weekly arousal to the normal pattern of a fluctuating T_b without arousal, or from a fluctuating T_b to a constant one.

Cheirogaleus medius seem to make the best of their tropical situation by suspending endogenous thermogenesis and accommodating ambient temperatures as ectothermic reptiles do. They continue to hibernate even at, or rather because of, temperatures above 30 °C. By minimizing the difference between T_b and T_h , they reduce their metabolism and energy expenditure while avoiding the energetic costs of arousals. This makes hibernation efficient, despite the high T_b attained each day.

To our knowledge, our findings are the first physiological confirmation of prolonged hibernation by a tropical mammal

Table 1 Comparison of heating phases in hibernating lemurs

	<i>N</i>	<i>n</i>	Amplitude (°C)	Heating time (h:min)	Heating rate (°C per h)
T_h	24	1,003	12.4 ± 4.1	$8:49 \pm 0:52$	1.4 ± 0.6
T_b	30	1,423	12.5 ± 3.3	$9:09 \pm 0:44$	1.4 ± 0.4
T_b^*	5	25	12.0 ± 1.8	$3:42 \pm 0:43^\dagger$	$3.4 \pm 1.0^\dagger$
χ^2			0.22	15.34	13.08
<i>P</i>			> 0.05	< 0.01	< 0.01

Features of heating phases are shown with daily fluctuations, measured from daily minima to daily maxima during the hibernation period, for tree-hole temperature (T_h) and body temperature (T_b) of *Cheirogaleus medius*, both in poorly insulated tree holes, and for the body temperature of *C. medius* (T_b^*) in well insulated tree holes during arousal. *N*, number of different animals tested; *n*, number of data points analysed, pooled for all animals tested.

[†]Post-hoc multiple comparisons following Kruskal–Wallis test showed T_b^* differed significantly from T_h and T_b .

as well as the first proof of hibernation in a primate. They show that hibernation can occur at a comparatively high T_b , so although previously defined as a state of hypometabolism and low T_b , the term 'hibernation' should now relate exclusively to hypometabolic states, irrespective of body temperature.

Kathrin H. Dausmann*, Julian Glos†, Jörg U. Ganzhorn‡, Gerhard Heldmaier*

*Department of Animal Physiology, Phillips University, 35043 Marburg, Germany e-mail: dausmann@staff.uni-marburg.de

†Department of Animal Ecology and Tropical Biology, Julius-Maximilians-University, 97074 Würzburg, Germany

‡Department of Animal Ecology and Conservation, University of Hamburg, 20146 Hamburg, Germany

1. Arnold, W. et al. *J. Therm. Biol.* **16**, 223–226 (1991).

2. Schmid, J. *Int. J. Primatol.* **19**, 797–808 (1998).

3. Wang, L. C. H. *Can. J. Zool.* **57**, 149–155 (1979).

4. Heldmaier, G., Ortman, S. & Körtner, G. in *Life in the Cold: Ecological, Physiological and Molecular Mechanisms* (eds Carey, C. et al.) 175–183 (Westview, Boulder, Colorado, 1993).

5. Humphries, M. M., Thomas, D. W. & Kramer, D. L. *Physiol. Biochem. Zool.* **76**, 165–179 (2003).

6. Carey, H. V., Andrews, M. T. & Martin, S. L. *Physiol. Rev.* **83**, 1153–1181 (2003).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Plant biochemistry

A naturally decaffeinated arabica coffee

The adverse side effects of caffeine¹ have increased the market for decaffeinated coffee to about 10% of coffee consumption worldwide (www.ncausa.org), despite the loss of key flavour compounds in the industrial decaffeinating process. We have discovered a naturally decaffeinated *Coffea arabica* plant from Ethiopia, a species normally recognized for the high quality of its beans. It should be possible to transfer this trait to commercial varieties of arabica coffee plants by intraspecific hybridization — a process likely to be simpler than an interspecific hybridization strategy, which could require more than 30 years of breeding to fix the decaffeinated trait and would probably result in an inferior cup of coffee.

Coffea arabica is the most cultivated and consumed coffee in the world. Attempts to transfer the caffeine-free property from wild coffee species of Madagascar, which produce an inferior beverage, to *C. arabica* have failed owing to a strong genetic barrier^{2,3}: *C. arabica* is an autogamous (over 95% self-pollinating) allotetraploid ($2n=4\times=44$), whereas all other coffee species are diploid and allogamous (none is self-pollinating).

Biotechnological attempts have also been made to interfere with caffeine production by preventing the expression of genes that encode key enzymes in the caffeine biosynthesis pathway⁴, culminating in the production of a transgenic *Coffea canephora* with

a 50–70% reduction of caffeine in the leaves⁵.

As part of a genetic breeding programme to reduce caffeine, initiated in 1987 at the Instituto Agrônomo de Campinas, we have studied 3,000 coffee trees, representing 300 *C. arabica* accessions from Ethiopia^{6,7}. We found that three of these Ethiopian plants (designated here as AC1, AC2 and AC3 in honour of the geneticist Alcides Carvalho) were almost completely free of caffeine. They were detected during routine high-performance liquid chromatography (HPLC) of methanolic extracts of seeds. The seeds of these plants showed a mean caffeine content of 0.76 mg g⁻¹ dry weight; Mundo Novo (MN), a commercial cultivar of *C. arabica*, contains 12 mg g⁻¹. The leaves and other parts of the fruit were also very low in caffeine.

We found that AC plants accumulated theobromine (about 6.1 mg g⁻¹ dry weight; Fig. 1a), the immediate precursor of caffeine⁸, indicating that these plants might be deficient in the enzyme caffeine synthase, which acts on theobromine. To test this, we infiltrated AC1 and MN leaves under vacuum with ¹⁴C-caffeine or ¹⁴C-adenine to monitor their degradation or their conversion to caffeine, respectively, in these plants. The radioactivity incorporated into metabolites was analysed after 4, 21 and 48 h of incubation by using HPLC coupled to ultraviolet and radioactivity detection systems⁹. (For methods, see supplementary information.)

The exogenous radiolabelled caffeine was degraded in both AC and MN plants (Fig. 1b). We found traces of radioactivity in theophylline, the first degradation product of caffeine, in both cases (results not shown). Theobromine can also occur as a degradation product of caffeine and although some radioactivity in theobromine was detected after treatment, this was minimal in both cases (Fig. 1b). These results indicate that the naturally low caffeine content in AC1 plants relative to MN plants is unlikely to be due to enhanced degradation of caffeine.

AC1 leaves accumulated radioactivity in theobromine when ¹⁴C-labelled adenine was fed to the plant, with none being incorporated into caffeine (Fig. 1c). By contrast, the amount of labelled theobromine in MN leaves decreased as the label became incorporated into caffeine (Fig. 1c). Because no caffeine synthase activity is detectable in AC1 leaves⁹, we conclude that the caffeine synthase gene is mutated in these AC plants.

Theobromine is the main alkaloid in cacao seeds (22–27 mg g⁻¹ dry weight) and has diuretic properties¹⁰; its effects may be similar to but much smaller than those of caffeine¹. However, its effects in AC1 coffee remain to be investigated. Information about cup quality and plant productivity is not yet available for AC plants but, given that *C. arabica* has a narrow genetic diversity and that even accessions from Ethiopia and Arabia (now Yemen), as well as old varieties,

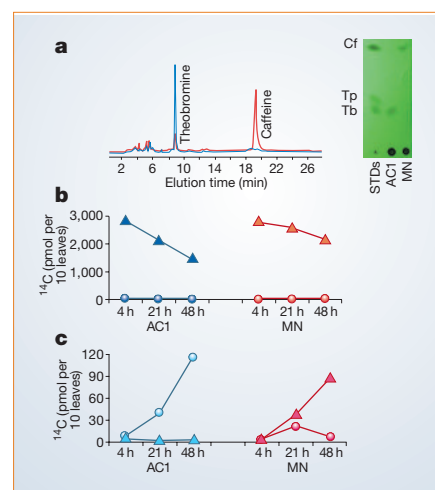


Figure 1 Caffeine production by *Coffea arabica* plants from Alcides Carvalho (AC1, red) and Mundo Novo (MN, blue). **a**, High-performance liquid chromatography (left) and thin-layer chromatography (right) profiles of methanolic extracts from leaves of AC1 and MN cultivars. Cf, caffeine; Tp, theophylline; Tb, theobromine; STDs, standards. **b**, **c**, Metabolism of exogenous ¹⁴C-caffeine (**b**) or ¹⁴C-adenine (**c**) in leaves of AC1 and MN plants to theobromine (circles) and caffeine (triangles).

all produce high-quality coffee¹¹, it is likely that AC plants will produce a good beverage.

AC plants all belong to the same accession, represented by 12 plants, which suggests that there has been segregation for the low-caffeine trait. We believe that by using a conventional breeding approach this trait could be successfully transferred to highly productive cultivars of *C. arabica*, because hybridizations will occur inside the same autogamous species. Our results highlight the value of maintaining and characterizing field germplasm collections.

Maria B. Silvarolla*, Paulo Mazzafera†, Luiz C. Fazuoli*

*Centro de Análise e Pesquisa Tecnológica do Agronegócio do Café 'Alcides Carvalho', IAC-APTA, CP 28, CEP 13001-970, Campinas, SP, Brasil

†Departamento de Fisiologia Vegetal, Instituto de Biologia, Universidade Estadual de Campinas, CP 6109, 13083-862, Campinas, Brasil

e-mail: pmazza@unicamp.br

1. James, J. E. *Caffeine and Health* (Academic, San Diego, 1991).

2. Charrier, A. *La Structure Génétique des Cafés Spontanés de la Région Malgache (Mascaraocoffea). Leurs Relations avec les Cafés d'Origine Africaine (Eucoffea)* (Orstom, Paris, 1978).

3. Mazzafera, P. & Carvalho, A. *Euphytica* **59**, 55–60 (1992).

4. Kato, M. & Mizuno, K. *Front. Biosci.* **9**, 1833–1842 (2004).

5. Ogita, S., Uefuji, H., Yamaguchi, Y., Koizumi, N. & Sano, H. *Nature* **423**, 823–823 (2003).

6. FAO Report: *Coffee mission to Ethiopia 1964–65* (Food and Agriculture Organization, Rome, 1968).

7. Silvarolla, M. B., Mazzafera, P. & Lima, M. M. A. *Gen. Mol. Biol.* **23**, 213–215 (2000).

8. Ashihara, H. & Suzuki, T. *Front. Biosci.* **9**, 1864–1876 (2004).

9. Mazzafera, P., Crozier, A. & Sandberg, G. *J. Agric. Food Chem.* **42**, 1423–1427 (1994).

10. Evans, W. C. *Trease and Evans's Pharmacognosy* (Saunders, London, 1996).

11. van der Vossen, H. A. M. in *Coffee — Botany, Biochemistry and Production of Beans and Beverage* (eds Clifford, M. N. & Wilson, K. C.) 48–107 (Helm, London, 1985).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Confronting the coral reef crisis

D. R. Bellwood¹, T. P. Hughes^{1,2}, C. Folke^{3,4} & M. Nyström³

¹Centre for Coral Reef Biodiversity, Dept. of Marine Biology, James Cook University, Townsville, Queensland 4811, Australia

²Ecole Pratique des Hautes Etudes, UMR CNRS 8046, Université de Perpignan, 66860 Perpignan Cedex, France

³Department of Systems Ecology, Stockholm University, SE-106 91 Stockholm, Sweden

⁴ Beijer International Institute of Ecological Economics, Royal Swedish Academy of Sciences, Stockholm, Sweden

The worldwide decline of coral reefs calls for an urgent reassessment of current management practices. Confronting large-scale crises requires a major scaling-up of management efforts based on an improved understanding of the ecological processes that underlie reef resilience. Managing for improved resilience, incorporating the role of human activity in shaping ecosystems, provides a basis for coping with uncertainty, future changes and ecological surprises. Here we review the ecological roles of critical functional groups (for both corals and reef fishes) that are fundamental to understanding resilience and avoiding phase shifts from coral dominance to less desirable, degraded ecosystems. We identify striking biogeographic differences in the species richness and composition of functional groups, which highlight the vulnerability of Caribbean reef ecosystems. These findings have profound implications for restoration of degraded reefs, management of fisheries, and the focus on marine protected areas and biodiversity hotspots as priorities for conservation.

The overall goal of coral reef management is to sustain the ability of tropical reefs to provide the ecosystem goods and services (for example, fisheries, tourism, aesthetic and cultural values), upon which human welfare depends¹. Although there have been some local successes, current management of reefs has failed to achieve this goal at a regional or global scale. Instead, coral reefs worldwide are in serious decline, owing primarily to over-harvesting^{2,3}, pollution^{4,5}, disease⁶ and climate change^{7–9}. Even the Great Barrier Reef, widely regarded as one of the most 'pristine' coral reefs in the world, shows system-wide decline (Fig. 1). In many locations around the world, man-made stresses to coral reefs have exceeded their regenerative capacity, causing dramatic shifts in species composition and resulting in severe economic loss.

In a changing world, one must expect and learn to manage

uncertainty¹⁰. In this review, we argue that the acceleration of human impacts on reef ecosystems requires a radical reassessment of the way these important marine resources are monitored and managed. We propose that the resilience of reef ecosystems—that is, their ability to absorb shocks, resist phase shifts and regenerate after natural and human-induced disturbances¹¹—needs to be more directly assessed and actively managed. Achieving this outcome requires an improved understanding of the dynamics of coral reef ecosystems, of the processes that support or undermine resilience, and of the socio-economic drivers and governance systems that regulate water quality and rates of extraction of reef resources. Here, we begin this reassessment by focusing on one important aspect of resilience, the role of critical functional groups^{12,13}. In this context, we document profound temporal and geographic variation in the ability of coral reefs to cope with accelerating human impacts.

Loss of resilience

Coral reefs, by definition, are three-dimensional shallow-water structures dominated by scleractinian corals. In the absence of severe human impacts, reefs readily reassemble after routine disturbances such as tropical hurricanes¹⁴. However, many contemporary coral reefs increasingly fail to regenerate after natural and human impacts, and instead have undergone a rapid shift to an alternate state^{15–18}. The most familiar of these transitions is from dominance by corals to dominance by fleshy seaweed, although several other transitions have been documented (Fig. 2). The extent to which alternate states are stable or reversible is poorly understood¹⁹ and represents a major challenge for research and management of reefs.

Until now there has been little success in predicting such regime or phase shifts, because the increased instability of coral reef ecosystems before their collapse has often been unrecognized, even on reefs that are relatively well studied. This cryptic loss of coral reef resilience can be manifested in numerous ways. For example, the collapse of many Caribbean coral reefs was long preceded by dwindling stocks of fishes and increased nutrient and sediment runoff from land^{2,16}. By the 1950s, when modern studies of reef ecology began, the prevention of macroalgal blooms was increasingly due to a single species of sea urchin, *Diadema antillarum*. In the 1970s, recorded densities of *Diadema* on overfished reefs were extraordinarily high, averaging more than ten individuals per square metre in shallow waters^{20–23}. The magnitude and crowded conditions of *Diadema* populations²⁴ may have contrib-

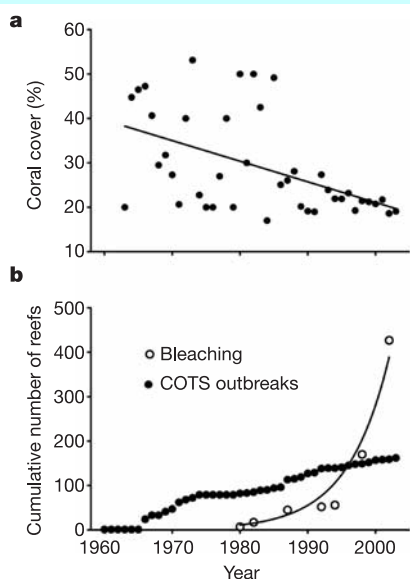


Figure 1 Degradation of coral reefs. **a**, Results of a meta-analysis of the literature, showing a decline in coral cover on the Great Barrier Reef. Each point represents the mean cover of up to 241 reefs sampled in each year. **b**, The recorded number of reefs on the Great Barrier Reef, Australia, substantially damaged over the past 40 yr by outbreaks of crown-of-thorns starfish (COTS) and episodes of coral bleaching.

uted to their eventual demise in 1983/4, when a disease outbreak spread throughout the Caribbean, reducing their numbers by two orders of magnitude²⁵ and precipitating macro-algal blooms that still persist. Today, remnant coral populations are further affected by increasingly prevalent coral disease and climatically induced coral bleaching^{6,9}.

With the benefit of hindsight, it is clear that long before the widespread loss of coral cover, many Caribbean reefs were on an unrecognized trajectory to collapse. The ecological symptoms included loss of macro-fauna^{2,16}, reduced fish stocks²⁶, a shift from fish-dominated to echinoid-dominated herbivory as the ecological role of herbivorous fishes was increasingly replaced by a single species of echinoid²², destructive overgrazing and bioerosion by food-limited sea urchins^{24,27}, and reduced coral recruitment²¹. Yet, although all of these features were exceptionally well documented, nobody put the pieces together in time to forecast their eventual consequence. We need to do better at recognizing and responding to these warnings.

A similar sequence of events is occurring on the Great Barrier Reef, where terrestrial runoff, over-harvesting and climate change are changing the dynamics and stability of the region²⁻⁵. Inputs of sediment and nutrients from land have increased fourfold since European settlement^{4,5}, while the numbers of turtles, dugongs and other macrofauna have greatly decreased^{28,29}. Modern management of the Great Barrier Reef began in 1975 with the establishment of the Great Barrier Reef Marine Park Authority, which protected 5% of the park from fishing. Comparisons of adjacent reefs open and closed to fishing today indicate that the biomass of targeted reef

fishes has been reduced by up to 60%, causing substantial changes in the abundance of their prey³⁰. Coral cover has significantly declined over the last 40 years (Fig. 1a), reflecting the impacts of three successive major outbreaks of crown-of-thorns starfish since the 1960s and two large-scale bleaching events in 1998 and 2000. In 2003, more than half the reefs sampled had <10% cover³¹. The low coral cover is likely to reflect marked demographic changes, reduced reproductive output of brood stocks, lower rates of recruitment, impaired connectivity, and species-level changes in coral composition (for example, in favour of short-lived, weedy taxa that recolonize more quickly). All of these dynamic features contribute to increased instability, yet none of them has been systematically tracked.

A functional group approach to coral reef dynamics

There are striking and profoundly important regional differences in the species richness, functional composition, dynamics and resilience of reef systems (Fig. 3). For example, Caribbean reefs have only a fraction of the number of species found on the Great Barrier Reef, approximately 28% for fishes and 14% for corals. We define a functional group as a collection of species that perform a similar function, irrespective of their taxonomic affinities¹². Although the Caribbean and Great Barrier Reef broadly share the same suite of functional groups, the species richness and taxonomic composition among functional groups is markedly different in the two regions. This difference is largely a biogeographic legacy of the evolutionary history of isolation and loss of taxa in the Caribbean basin^{32,33}. The result is a functionally compromised assemblage that is more

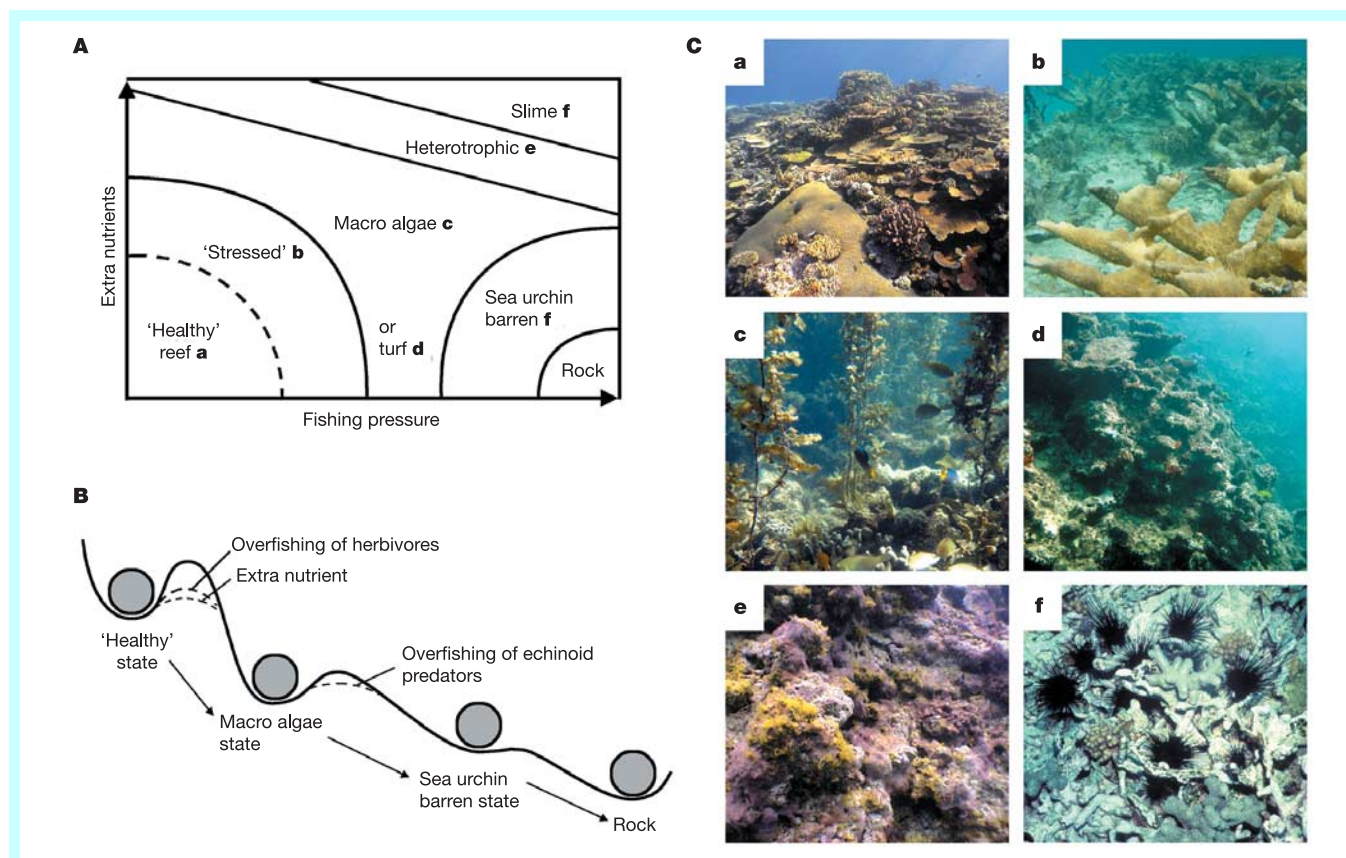


Figure 2 Alternate states in coral reef ecosystems. **A**, A conceptual model showing human-induced transitions between alternate ecosystem states based on empirical evidence of the effects from fishing and excess nutrients¹⁵⁻¹⁷. The 'stressed' state illustrates loss of resilience and increased vulnerability to phase-shifts. **B**, A graphic model depicting transitions between ecosystem states. 'Healthy' resilient coral-

dominated reefs become progressively more vulnerable owing to fishing pressure, pollution, disease and coral bleaching. The dotted lines illustrate the loss of resilience that becomes evident when reefs fail to recover from disturbance and slide into less desirable states. **C**, Six characteristic reef states (as in **A**) from sites on the Great Barrier Reef (**a**, **c**, **d**, **e**) and in the Caribbean (**b**, **f**).

vulnerable to catastrophic phase shifts, particularly when subject to human exploitation and impacts.

Fish functional groups are generally synonymous with guilds of species from different trophic levels within a food chain (for example, predators and herbivores)³⁴, reflecting their role as a major conduit for the flow of energy on reefs. In this study, however, functional groupings are also identified by their roles in ecosystem processes. In contrast to fishes, the greatest contribution of corals is in the accumulation of carbonate and provision of structure. For corals, we can classify functional groups on the basis of the shape of their colonies, reflecting their role in reef processes. These include the creation of three-dimensional habitats for fishes and other organisms and their contribution to reef growth as either primary or secondary framework builders³⁴. Corals, like fishes, often play multiple functional roles and thus support several different reef processes. These functional groups also have strikingly divergent ecologies and life histories, and exhibit marked differences in their susceptibility to hurricanes, sedimentation and other disturbances³⁵.

In the Caribbean, several critical functional groups are missing or represented by only a handful of species (Fig. 3). There are, for example, no three-dimensional bottlebrush species and just one staghorn (*Acropora cervicornis*) and one tall, tabular coral (*Acropora palmata*). Importantly, these are the dominant habitat-creating functional groups on healthy reefs in both the Indo-Pacific and Caribbean. Until recently, the two Caribbean species were abundant and widespread, commonly comprising more than 30–50% of the total coral cover down to a depth^{36–38} of 20 m. Today, many areas have effectively lost not only these two species, but also two critical functional groups and two major shallow-water reef habitats: the

elkhorn ‘*palmata* zone’ and the staghorn ‘*cervicornis* zone’³⁶. Both species were added in 1999 to the Candidate Species List of the US Endangered Species Act³⁷, a sad reflection of our inability to implement regional-scale management of Caribbean reefs.

For fishes, the composition of functional groups on Caribbean reefs is also markedly different from that of the Great Barrier Reef (Fig. 3). Nocturnal and diurnal planktivores, in particular, are greatly under-represented in the Caribbean. For both fishes and corals, the Great Barrier Reef fauna has more species than the Caribbean in all functional groups (Fig. 3). This may confer a higher degree of functional redundancy within groups on the Great Barrier Reef, where the loss of any one species is potentially compensated for by the actions of another.

But does high species richness confer a degree of ecological insurance for ecosystem performance, as suggested by some small-scale experimental studies of biodiversity and ecosystem function³⁹? The available evidence for coral reefs is equivocal. High diversity undoubtedly provides the potential for functional redundancy. However, even in high-diversity systems redundancy in critical functional groups can be limited⁴⁰. Conversely, low diversity reefs at some isolated locations, such as Clipperton Atoll, survive with minimal representation in major functional groups⁴¹.

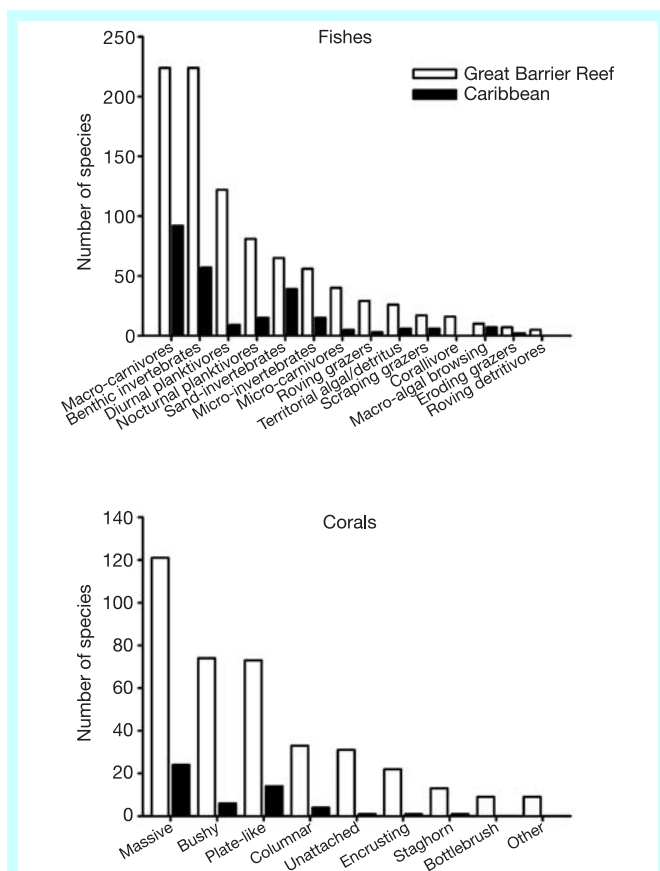


Figure 3 Functional composition of Caribbean and Great Barrier Reef assemblages of fishes and corals. The fourteen fish and eleven coral functional groups are identified by their roles in ecosystem processes.

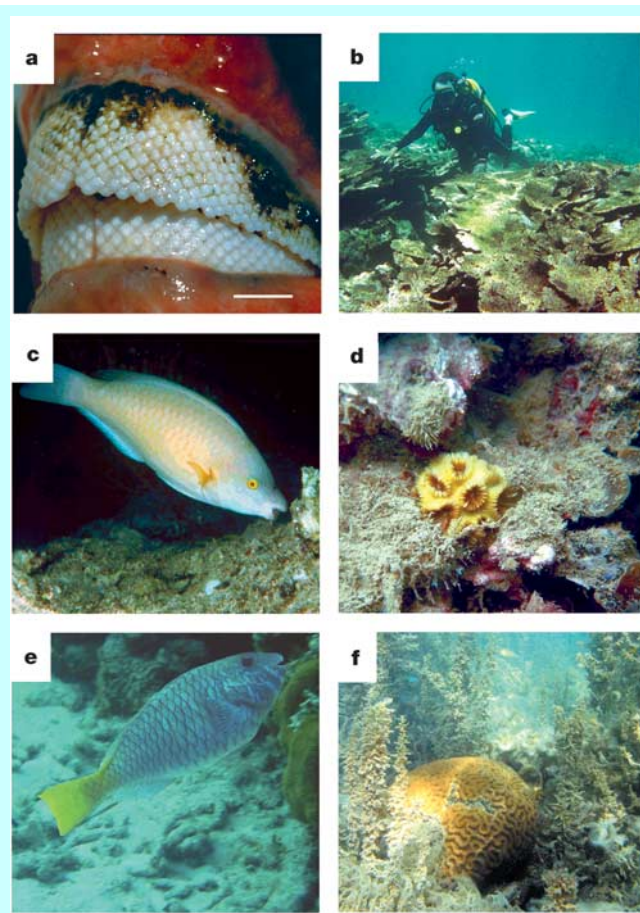


Figure 4 Three critical functional groups and their roles in facilitating reef recovery. **a**, The jaws of a bioeroding parrotfish (*Bolbometopon muricatum*); each individual ingests five tonnes of coral annually. Scale 1 cm. **b**, An extensive stand of dead coral in Samoa, killed by coral bleaching, remains intact because of reduced bioerosion by depleted fish populations. **c**, A scraping herbivore, the parrotfish (*Scarus flavipectoralis*) removes epilithic algae and sediment. **d**, A juvenile coral overwhelmed by algae and trapped sediment. **e**, A grazing parrotfish (*Sparisoma rubripinne*) reduces overgrowth of corals by competing macroalgae. **f**, An adult coral shaded and overgrown by fleshy macro-algae.

Nevertheless, the clearest examples available emphasize the impact of limited redundancy in relatively depauperate locations. For example, sea urchins became the principal herbivores on many Caribbean reefs following the depletion of fishes, and they prevented the rapid phase shift to dominance by macro-algae that was precipitated by the die-off of *Diadema*^{16,17,23}.

However, the changes to Caribbean reefs also demonstrate that the loss of functional redundancy can come at great cost, even when some members of a group compensate for others. Although fishes

and echinoids both consume algae, and can substitute for each other in this role, echinoids are far more destructive bioeroders. Only a few species of parrotfish erode significant volumes of reef carbonate when feeding, and they feed primarily on dead corals and other protuberances, avoiding flat surfaces^{33,42}. In marked contrast, grazing echinoids burrow into and erode the reef matrix²⁷. In high densities they can undercut and dislodge massive corals. If unchecked, urchins have the capacity to destroy reefs, as documented in the Galapagos Islands and elsewhere in the East Pacific^{43,44}, where the reef structure has been eroded at rates of up to $10 \text{ kg m}^{-2} \text{ yr}^{-1}$.

Furthermore, resilience is critically dependent on the range of responses to environmental change by species within each functional group—that is, their response diversity⁴⁵. Clearly, functional redundancy is ineffective if every species comprising a functional group always reacts to a disturbance in the same way and to the same extent. For example, the response of reef fishes to chronic overfishing leaves few species intact. In such cases, the insurance value of redundancy and of high species richness is negligible.

Resilience and functional groups

Three functional groups (conventionally combined as ‘herbivores’ and dominated by fishes), play different and complementary roles in preconditioning reefs to permit recovery of corals. These three groups—bioeroders, scrapers and grazers—are a critical source of both resilience and vulnerability to phase shifts. Bioeroding fishes remove dead corals^{40,42}, exposing the hard, reef matrix for settlement of coralline algae and corals. Scrapers directly remove algae and sediment by close cropping, facilitating settlement, growth and survival of coralline algae and corals^{46,47}. Grazers remove seaweed, reducing coral overgrowth and shading by macro-algae^{16,48} (Fig. 4).

Box 1

Functional groups, fisheries and economic development

Overfishing is a major environmental and economic problem facing virtually all marine ecosystems^{2,71}. Typically, overexploitation of a mixed reef fishery first depletes stocks of large predators (for example, sharks and groupers), and herbivorous fishes and planktivores subsequently become a more prevalent component of the total catch^{26,71}. Consequently, it is increasingly difficult to evaluate the ecological effect of loss of predators because few places remain with relatively intact fish faunas to serve as experimental controls⁷². Nonetheless, comparisons of lightly and heavily fished sites provide evidence for top-down alterations to food webs after depletions of predators, on the Great Barrier Reef³⁰ and in Fiji⁷³. These trophic cascades occur despite the potential functional redundancy of fish predators, which are represented by approximately 18 families and up to 200 species in a typical central Indo-Pacific reef system (Fig. 3). Importantly, in more species-impooverished bioregions, lower functional redundancy and response diversity may lead to substantially greater impacts from loss of predators. In parts of the Caribbean and Eastern Pacific, for example, depletion of fish predators of echinoids is likely to have played a key role in generating unsustainably high densities of sea urchins²². Similarly, the widespread declines of herbivorous and predatory turtles are likely to have increased the biomass of seagrasses and sponges².

Until recently, fishing on most coral reefs has been largely artisanal, providing a much-needed and cheap source of protein. Nevertheless, in many Pacific locations, traditional fishing has severely undermined major components of the bioeroding group and compromised ecosystem function⁴⁰. Similarly, in Jamaica and elsewhere in the Caribbean, intense artisanal fishing focuses heavily on grazers²⁶. In recent decades, however, there has been a dramatic increase in fishing effort on coral reefs, and export of both live and dead coral reef fishes is expanding rapidly. Parrotfishes from the Seychelles and Persian Gulf are now sold by fish retailers in London. The unprecedented demand for live reef fishes in China, Singapore and Taiwan is financed by the expanding disposable incomes of a large Asian hinterland, and exerts additional fishing pressure on reefs throughout vast areas of the Indo-Pacific^{74,75}. With retail prices⁷⁶ of up to US\$250 per kg, exploitation of remote reef systems has become financially viable for the first time, overcoming previous cost barriers. Herbivorous fishes are an increasingly significant component of the live fish trade, ranking currently as the second largest group targeted for exploitation (based on biomass) (Fig. 5). These new markets have greatly augmented both the intensity and scale of exploitation, and are set to increase as fish stocks elsewhere continue to decline⁷¹. The depletion of herbivorous fishes combined with increasingly frequent bleaching events is an ominous combination, but one faced by numerous reef systems around the globe⁸.

The exploitation of reef fish, such as groupers and parrotfishes, clearly illustrates a mismatch between the global demand for reef fishes and the fundamental role of functional groups in ecosystem resilience. Phase-shifts of tropical reefs to less desirable states (Fig. 2) can have devastating economic effects on maritime developing nations¹. Without healthy reef systems, future options for social and economic development will be constrained or lost.

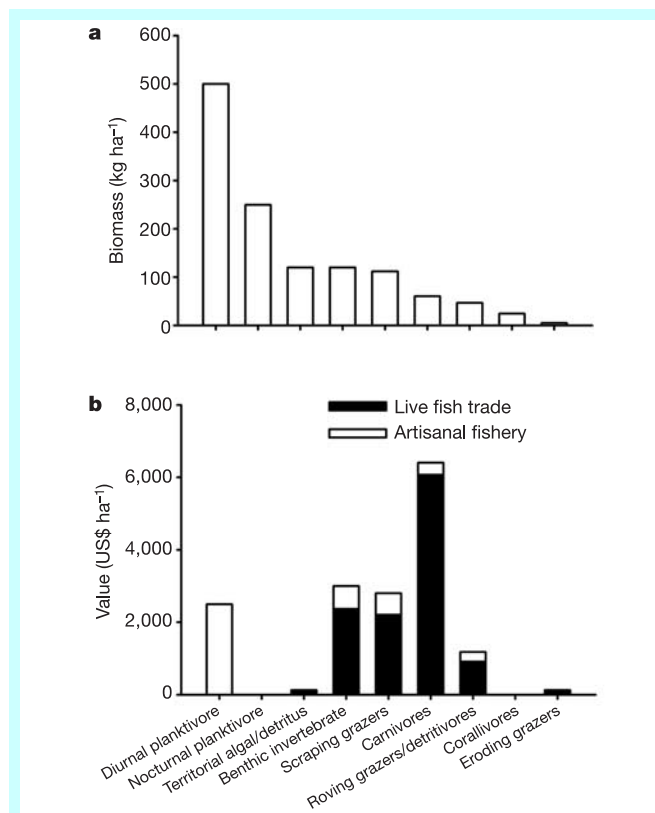


Figure 5 The mismatch between the economics of consumption and the ecological role of fishes on coral reefs. **a**, **b**, The biomass of functional groups of reef fishes in a central Indo-Pacific reef system (**a**), compared to the economic value of the fishes to an artisanal fishery or the international live fish trade (**b**).

The extent to which reefs possess these functional groups is central to their capacity to resist phase shifts, regenerate and retain critical functions in the face of disturbance.

A key element in resisting phase shifts to degraded alternate states is the maintenance of successful larval colonization by the full range of coral functional groups characteristic of the region. (Regrowth of residual corals can also be locally important, but they too must be replaced ultimately by new recruits). Degraded reefs exhibit declining rates of coral recruitment, through a poorly understood combination of reduced adult fecundities, lower settlement, and higher rates of early mortality (for example, due to smothering by algae and sediment)^{46,49–51}. Reduced rates of fish bioerosion and increased mortality of corals have dramatically increased the prevalence of a relatively new phenomenon: large tracts of intact dead coral skeletons⁷. Coral settlement on such physically fragile or unstable foundations results in abbreviated recovery and a shift to weedy coral species. The scale of dispersal of coral larvae is also crucial to understanding coral reef reorganization after large-scale disturbances. The degree of connectivity and gene flow between local populations varies substantially among corals: some settle quickly and are largely self-seeded; others are more widely dispersed⁵². The isolation of oceanic reefs renders them particularly vulnerable to loss of local broodstocks⁵³. On degraded reefs, the local loss of brood stocks is likely to select against self-seeding species and to shift the taxonomic composition of recruits in favour of those with longer planktonic durations, with potentially far-reaching consequences for community structure. Even if local conditions for settlement improve (for example, as a result of reduced overfishing, or a partial recovery of *Diadema antillarum*), it can no longer be assumed that coral recruitment will simply resume exactly as before.

Regenerative capacities need to be better understood and actively managed, so that human beings can become a more efficient and much less destructive component of coral reef ecosystems. If coral reefs are to resist phase shifts after disturbance it is imperative that critical functional groups of fishes, corals and other taxa are actively managed and sustained. Unfortunately, the critical groups that underpin the formation of three-dimensional coral growth and herbivory by fishes are increasingly threatened at precisely the time when the impacts of human disturbance to coral reefs make their functions as promoters of coral reef resilience all the more essential^{8,11} (Box 1).

Shifting baselines and adaptive management

It is increasingly clear that the rapid decline of reef systems calls for a suite of more vigorous, innovative and adaptive management strategies. Responding to the global coral reef crisis requires active management of human activities that modify essential ecological processes. In particular, it requires an ability to scale up management and governance systems to secure the future of functional groups and their roles in supporting the resilience of coral reefs. There is a growing awareness of what has already been lost^{2,3}, and also a recognition that in a changing world, the resilience of coral reefs is increasingly uncertain⁸.

Management of tropical fisheries, where it exists, has almost always been instigated long after exploitation has peaked, with the goal of sustainably harvesting whatever little remains. Typically, the stocks continue to decline even further, and over time management targets slip lower and lower, a scenario known as “the shifting baseline”⁵⁴. Today, for example, a new generation of Caribbean researchers and managers may never have seen a decent stand of Caribbean *Acropora* coral, a manatee or a large shark, nor can they remember the destruction wrought in the 1970s by a million sea urchins per kilometre of coastline. Shifting baselines such as these pervade coral reef science and management⁵⁵. Recent descriptions of purported recovery of degraded reefs in the Caribbean and Hawaii^{51,56}, and well-meaning efforts to culture and reintroduce *Diadema* dramatically illustrate the issue. Similarly, there is a

common tendency to attribute reef condition solely to more recent and better-studied impacts, underplaying the importance of historical disturbances^{2,3,8}. Much has been lost, and some of it forever.

Management of functional groups represents a radical departure from current management philosophy. The fundamental difference is that management of functional groups recognizes that the cost of failure extends beyond the immediate impacts of depletion of over-exploited fish stocks or reductions in coral cover. Thus, the management of herbivorous fishes can facilitate the regeneration of reefs after large-scale disturbances such as bouts of bleaching or disease that are impossible to regulate locally. Critically, a functional approach provides the basis for managing uncertainty by maintaining the functional groups that support dynamic ecological processes (for example, herbivory and provision of habitat), in contrast to the conventional goals of maintaining the status quo (high coral cover and sustainable fisheries yields).

Today, ‘no-take’ areas (NTAs), where fishing and other human activities are prohibited, are an increasingly prevalent approach to coral reef management^{55,57,58}. If they are adequately enforced, NTAs provide a spatial refuge from harvesting. Importantly, such protection may also permit critical functional groups to persist, and thus contribute to local ecosystem resilience. However, even the largest NTAs in the world are not self-sustaining, because they are too small relative to the scale of natural and human disturbances, and to the dispersal distances of many larvae and migrating adults⁵⁹. Currently, most NTAs are a few square kilometres or less in size, and they are invariably surrounded by vastly larger areas that are often already badly degraded. Although one of the main benefits of NTAs is the export of propagules and adults, on coral reefs the initial success of existing NTAs can often be attributed to an influx of larvae⁵⁷. It is crucial, therefore, that NTAs are viewed more realistically in the context of the whole seascape. There is a distinct danger in overselling the benefits of establishing a few highly protected areas at specific locations, whose purported role is to remain or once more become “pristine”⁶⁰. Indeed, as human impacts continue, recruitment from degrading reefs into NTAs is not only likely to decline but also may include a growing component of undesirable species (for example, algal spores, pathogens and introduced species). Ultimately, the long-term success of NTAs and the status of surrounding areas should be evaluated in terms of the processes and mechanisms that contribute to the resilience of reefs, not just estimates of abundance or counts of selected species. This caveat applies equally to another current issue for contemporary management of coral reefs—the focus on biodiversity hotspots.

Hotspots, areas of exceptional species richness⁶¹, are one of the most frequently identified targets for the protection of marine ecosystems⁶². However, there are several lines of evidence to suggest that ‘cool spots’, areas of low species richness, are more vulnerable^{63–65}. Low-diversity reefs, such as in the Caribbean Basin, the Eastern Pacific, and many high-latitude or remote locations in the Indo-Pacific have low functional redundancy, where functional groups may be represented by a single species. In these systems, as noted above, minor changes in biodiversity can have a major impact on ecosystem processes and consequently on the people whose livelihoods depend on the services that ecosystems generate.

A blueprint for the future

We conclude with four major recommendations for managing human activities in coral reef ecosystems. First, the rate of establishment and size of NTAs, as a tool for resilience management, needs to be hugely increased. In Australia, the expansion of NTAs in 2004 from <5% to 33% of the Great Barrier Reef Marine Park, with a parallel focus on improving water quality, provides a good model⁶⁶. In the United States, in comparison, there are more modest plans⁶⁷ to increase NTAs to incorporate 20% of reefs by 2010, a clear case of too little, too late. Currently, these two affluent nations are responsible for over two-thirds of the world’s coral reef Protected Areas⁶⁸.

Elsewhere, developing countries are faced with a serious lack of resources, which limits the number, size and efficacy of NTAs and increases the likelihood of 'paper parks'. International efforts in support of marine parks for promoting resilience need to be substantially expanded. Second, the focus on NTAs and hotspots must not be allowed to detract from the provision of improved management measures for the vast majority of reefs that are heavily affected by people⁶⁹. Unless we can achieve regional-scale active management of critical functional groups to support reef resilience, any small-scale successes within NTAs or by individual countries may be unable to stem the decline of reef systems as a whole. Third, reef management needs to be more inclusive, proactive and responsive. Governance systems should support ownership and empowerment of users as stewards of reef resilience, provide incentives for herbivore protection before—not after—stocks collapse, and implement flexible restrictions (for example, to enhance the protection of critical broodstocks during the vulnerable spawning period). Fourth, markets for reef resources must be reformed to incorporate economic incentives that prevent exploitation of species in critical functional groups (Box 1). Such action is unlikely to emerge from uninformed human preferences. Markets for reef resources urgently need to be framed by norms and rules (institutions), operating from local to global scales, that secure coral reef resilience and thereby promote a greater diversity of options for economic development⁷⁰.

Developing new metrics for stewardship of coral reef resilience is vital for coping with uncertainty and surprise in a biosphere increasingly shaped by human action. In this context the focus, both within and outside NTAs, should shift from conservation of species to active management of critical functional groups that support important processes and sustain ecosystem services. Current aspirations towards sustaining fisheries need to fundamentally change their metrics from stock assessments to capturing the ecosystem performance and resilience that support long-term fisheries production⁷¹. Ecosystem metrics for monitoring the status of coral reefs need to move beyond coral cover and counts of targeted species to include functional groups, functional redundancy and response diversity. In all of these endeavours, we are faced with a critical lack of knowledge. Our ability to continue to exploit coral reef resources will depend on an effective combination of science-based management, public support and political will. Clearly, successful management of coral reef ecosystems will also require courage, creativity and a willingness to move beyond traditional metrics, models and perceptions. □

doi:10.1038/nature02691.

1. Moberg, F. & Folke, C. Ecological goods and services of coral reef ecosystems. *Ecol. Econ.* **29**, 215–233 (1999).
2. Jackson, J. B. C. *et al.* Historical overfishing and the recent collapse of coastal ecosystems. *Science* **293**, 629–638 (2001).
3. Pandolfi, J. M. *et al.* Global trajectories of the long-term decline of coral reef ecosystems. *Science* **301**, 955–958 (2003).
4. Williams, D., McB., *et al.* *The Current Level of Scientific Understanding on Impacts of Terrestrial Run-Off on the Great Barrier Reef World Heritage Area.* (http://www.reef.crc.org.au/aboutreef/coastal/waterquality_consensus.html) (2002).
5. McCulloch, M. *et al.* Coral record of increased sediment flux to the inner Great Barrier Reef since European settlement. *Nature* **421**, 727–730 (2003).
6. Harvell, C. D. *et al.* Climate warming and disease risks for terrestrial and marine biota. *Science* **296**, 2158–2162 (2002).
7. Wilkinson, C. (ed.) *Status of Coral Reefs of the World* (Australian Institute of Marine Science, Townsville, Australia, 2002).
8. Hughes, T. P. *et al.* Climate change, human impacts, and the resilience of coral reefs. *Science* **301**, 929–933 (2003).
9. Gardener, T. A., Cote, I., Gill, J. A., Grant, A. & Watkinson, A. R. Long-term region-wide declines in Caribbean corals. *Science* **301**, 958–960 (2003).
10. Kinzig, A. P. *et al.* Coping with uncertainty: A call for a new science-policy forum. *Ambio* **32**, 330–335 (2003).
11. Nyström, M., Folke, C. & Moberg, F. Coral reef disturbance and resilience in a human-dominated environment. *Trends Ecol. Evol.* **15**, 413–417 (2000).
12. Steneck, R. S. & Dethier, M. N. A functional group approach to the structure of algal-dominated communities. *Oikos* **69**, 476–498 (1994).
13. Walker, B. H. in *Plant Functional Types* (eds Smith, T. M., Shugart, H. H. & Woodward, F. I.) 91–103 (Cambridge Univ. Press, UK, 1997).
14. Connell, J. H., Hughes, T. P. & Wallace, C. C. A 30-years study of coral abundance, recruitment, and disturbance at several scales in space and time. *Ecol. Monogr.* **67**, 461–488 (1997).
15. Done, T. J. Phase shifts in coral reef communities and their ecological significance. *Hydrobiologia* **247**, 121–132 (1992).
16. Hughes, T. P. Catastrophes, phase-shifts, and large-scale degradation of a Caribbean coral reef. *Science* **265**, 1547–1551 (1994).
17. Knowlton, N. Thresholds and multiple stable states in coral reef community dynamics. *Am. Zool.* **32**, 674–682 (1992).
18. Scheffer, M. *et al.* Catastrophic shifts in ecosystems. *Nature* **413**, 591–596 (2001).
19. Scheffer, M. & Carpenter, S. R. Catastrophic regime shifts in ecosystems: linking theory to observation. *Trends Ecol. Evol.* **18**, 648–656 (2003).
20. Ogden, J. C., Brown, R. A. & Salesky, N. Grazing by the echinoid *Diadema antillarum* Philippi: Formation of halos around West Indian patch reefs. *Science* **182**, 715–717 (1973).
21. Sammarco, P. W. *Diadema* and its relationship to coral spat mortality: grazing, competition, and biological disturbance. *J. Exp. Mar. Biol. Ecol.* **45**, 245–272 (1980).
22. Hay, M. E. Patterns of fish and urchins grazing on Caribbean coral reefs: are previous results typical? *Ecology* **65**, 446–454 (1984).
23. Lessios, H. A. Mass mortality of *Diadema antillarum* in the Caribbean: what have we learned? *Annu. Rev. Ecol. Syst.* **19**, 371–393 (1988).
24. Levitan, D. R. Density-dependent size regulation and negative growth in the sea urchin *Diadema antillarum* Philippi. *Oecologia* **76**, 627–629 (1988).
25. Lessios, H. A., Robertson, D. R. & Cubitt, J. D. Spread of *Diadema* mass mortality through the Caribbean. *Science* **226**, 335–337 (1984).
26. Munro, J. L. (ed.) *Caribbean Coral Reef Fisheries* 2nd edn, Vol. 7, 1–276 (ICLARM Stud. Rev., ICLARM, Makati, 1983).
27. Ogden, J. C. in *Reefs and Related Carbonates—Ecology and Sedimentology* (eds Frost, S. H., Weiss, M. P. & Saunders, J. B.) Stud. Geol. **4** 281–288 (Am. Assoc. Petrol. Geol., USA, 1977).
28. Heppell, S. S. *et al.* Population model analysis for the loggerhead sea turtle, *Caretta caretta*, in Queensland. *Wildl. Res.* **23**, 143–159 (1996).
29. Great Barrier Reef Marine Park Authority. Overview: the current status of the Great Barrier Reef. (http://www.gbrmpa.gov.au/corp_site/info_services/publications/sotr/downloads/SPRR_Overview2.pdf) (2003; cited December 2003).
30. Graham, N. A. J., Evans, R. D. & Russ, G. R. The effects of marine reserve protection on the trophic relationships of reef fishes on the Great Barrier Reef. *Environ. Conserv.* **30**, 200–208 (2003).
31. AIMS. Australian Institute of Marine Science. *Reef Monitoring* <http://www.aims.gov.au/pages/research/reef-monitoring/reef-monitoring-index.html>. (2003; cited December 2003)
32. Johnson, K. G., Budd, A. F. & Steemann, T. A. Extinction selectivity and ecology of Neogene Caribbean reef corals. *Paleobiology* **21**, 52–73 (1995).
33. Bellwood, D. R. & Wainwright, P. C. in *Coral Reef Fishes: Dynamics and Diversity in a Complex Ecosystem* (ed. Sale, P. F.) 5–32 (Academic, San Diego, 2002).
34. Done, T. J., Ogden, J. C., Wiebe, W. J. & Rosen, B. R. in *Functional Roles of Biodiversity: a Global Perspective* (eds Mooney, J. H., Cushman, E., Medina, E., Sala, O. E. & Schulze, E.-D.) 393–429 (John Wiley, Sons, New York, 1996).
35. Hughes, T. P. & Connell, J. H. Multiple stressors on coral reefs: A long-term perspective. *Limnol. Oceanogr.* **44**, 932–940 (1999).
36. Goreau, T. F. The ecology of Jamaican coral reefs. 1. Species composition and zonation. *Ecology* **40**, 67–89 (1959).
37. NOAA Fisheries. Candidate Corals: *Acropora palmata* (Elkhorn Coral) and *Acropora cervicornis* (Staghorn Coral) (<http://www.nmfs.noaa.gov/habitat/ecosystem/AcroporaWorkshop.htm>).
38. Aronson, R. B., Macintyre, I. G., Precht, W. F., Murdoch, T. J. T. & Wapnick, C. M. The expanding scale of species turnover events on coral reefs in Belize. *Ecol. Monogr.* **72**, 233–249 (2002).
39. Loreau, M., Naeem, S. & Inchausti, P. (eds) *Biodiversity and Ecosystem Functioning: Synthesis and Perspectives* 1–306 (Oxford Univ. Press, Oxford, 2002).
40. Bellwood, D. R., Hoey, A. S. & Choat, J. H. Limited functional redundancy in high diversity systems: resilience and ecosystem function on coral reefs. *Ecol. Lett.* **6**, 281–285 (2003).
41. Glynn, P. W. & Ault, J. S. A biogeographic analysis and review of the far eastern Pacific coral reef region. *Coral Reefs* **19**, 1–23 (2000).
42. Bellwood, D. R. & Choat, J. H. A functional analysis of grazing in parrotfishes (family Scaridae): the ecological implications. *Environ. Biol. Fishes* **28**, 189–214 (1990).
43. Eakin, C. M. Where have all the carbonates gone? A model comparison of calcium carbonate budgets before and after the 1982–1983 El Niño at Uva Island in the eastern Pacific. *Coral Reefs* **15**, 109–119 (1996).
44. Glynn, P. W. El Niño warming, coral mortality and reef framework destruction by echinoid bioerosion in the eastern Pacific. *Galaxea* **7**, 129–160 (1988).
45. Elmqvist, T. *et al.* Response diversity, ecosystem change and resilience. *Front. Ecol.* **1**, 488–494 (2003).
46. Hunte, W. & Wittenberg, M. Effects of eutrophication and sedimentation on juvenile corals. II. Settlement. *Mar. Biol.* **114**, 625–631 (1992).
47. Steneck, R. S., in *6th International Coral Reef Symposium* (eds Choat, J. H. C. *et al.*) Vol. 1 37–49 (6th International Coral Reef Symposium Executive Committee, Townsville, 1988).
48. McCook, L. J., Jompa, J. & Diaz-Pulido, G. Competition between corals and algae on coral reefs: a review of available evidence and mechanisms. *Coral Reefs* **19**, 400–417 (2001).
49. Richmond, R. H. Coral reefs: present problems and future concerns resulting from anthropogenic disturbances. *Am. Zool.* **33**, 524–536 (1993).
50. Hughes, T. P. & Tanner, J. E. Recruitment failure, life histories, and long-term decline of Caribbean corals. *Ecology* **81**, 2250–2264 (2000).
51. Edmunds, P. J. & Carpenter, R. C. Recovery of *Diadema antillarum* reduces macroalgal cover and increases abundance of juvenile corals on a Caribbean reef. *Proc. Natl Acad. Sci. USA* **98**, 5067–5073 (2001).
52. Ayre, D. J. & Hughes, T. P. Genotypic diversity and gene flow in brooding and spawning corals along the Great Barrier Reef, Australia. *Evolution* **54**, 1590–1605 (2000).
53. Ayre, D. J. & Hughes, T. P. Climate change, genotypic diversity and gene flow in reef-building corals. *Ecol. Lett.* **7**, 273–278 (2004).
54. Pauly, D. Anecdotes and the shifting baseline syndrome of fisheries. *Trends Ecol. Evol.* **10**, 430 (1995).
55. Bohnsack, J. A. Shifting baselines, marine reserves, and Leopold's biotic ethic. *Gulf Caribb. Res.* **14**, 1–7 (2003).
56. Scully, E. P. & Ostrander, G. K. Corals and phase shifts. *Trends Ecol. Evol.* **16**, 126–127 (2001).

57. Russ, G. R. in *Coral Reef Fishes: Dynamics and Diversity in a Complex Ecosystem* (ed. Sale, P. F.) 421–443 (Academic, San Diego, 2002).
58. Halpern, B. S. & Warner, R. R. Marine reserves have rapid and lasting effects. *Ecol. Lett.* **5**, 361–366 (2002).
59. Nyström, M. & Folke, C. Spatial resilience of coral reefs. *Ecosystems* **4**, 406–417 (2001).
60. Folke, C., Holling, C. S. & Perrings, C. Biological diversity, ecosystems, and the human scale. *Ecol. Appl.* **4**, 1018–1024 (1996).
61. Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. & Kent, J. Biodiversity hotspots and conservation priorities. *Nature* **403**, 853–858 (2000).
62. Roberts, C. M. *et al.* Marine biodiversity hotspots and conservation priorities for tropical reefs. *Science* **295**, 1280–1284 (2002).
63. Bellwood, D. R. & Hughes, T. P. Regional-scale assembly rules and biodiversity of coral reefs. *Science* **292**, 1532–1534 (2001).
64. Hughes, T. P., Bellwood, D. R. & Connolly, S. R. Biodiversity hotspots, centers of endemism, and the conservation of coral reefs. *Ecol. Lett.* **5**, 775–784 (2002).
65. Kareiva, P. & Marvier, M. Conserving biodiversity coldspots. *Am. Sci.* **91**, 344–351 (2003).
66. Great Barrier Reef Marine Park Authority Representative Areas Program (http://www.reefed.edu.au/rap/pdf/rap_overview_brochure.pdf) (2003).
67. USCRF *The National Action Plan to Conserve Coral Reefs* (<http://coralreef.gov>) (US Coral Reef Task Force, Washington, DC, 25 December 2003).
68. Spalding, M. D., Ravilious, C. & Green, E. P. *World Atlas of Coral Reefs* (Univ. California Press, Berkeley, 2001).
69. Folke, C. *et al.* Regime shifts, resilience and biodiversity in ecosystem management. *Annu. Rev. Ecol. Syst.* (in the press).
70. Dietz, T., Ostrom, E. & Stern, P. C. The struggle to govern the commons. *Science* **302**, 1907–1912 (2003).
71. Pauly, D. *et al.* Towards sustainability in world fisheries. *Nature* **418**, 689–695 (2002).
72. Steneck, R. S. Human influences on coastal ecosystems: does overfishing create trophic cascades? *Trends Ecol. Evol.* **13**, 429–430 (1998).
73. Jennings, S. & Polunin, N. V. C. Impacts of predator depletion by fishing on the biomass and diversity of non-target reef fish communities. *Coral Reefs* **16**, 71–82 (1997).
74. Birkeland, C. Symbiosis, fisheries and economic development on coral reefs. *Trends Ecol. Evol.* **12**, 364–367 (1997).
75. Sadovy, Y. J. & Vincent, A. C. J. in *Coral Reef Fishes: Dynamics and Diversity in a Complex Ecosystem* (ed. Sale, P. F.) 391–420 (Academic, San Diego, 2002).
76. McGilvray, F. & Chan, T. T. C. The trade in live reef foodfish: a Hong Kong perspective. *Aquaculture Asia* **7**, 21–26 (2002).

Acknowledgements We thank the Australian Research Council, the Queensland Government, and the Beijer Institute of the Royal Swedish Academy of Sciences for facilitating meetings of the authors in Australia and Sweden; the Laboratoire de Biologie Marine et Malacologie at the University of Perpignan for hosting a sabbatical visit by T.P.H.; the Great Barrier Reef Marine Park Authority for provision of images; Lizard Island Research Station for field support; colleagues in the Resilience Alliance, Centre for Coral Reef Biodiversity, James Cook University, and Department of Systems Ecology, Stockholm University, for comments or discussions. This work was supported by the Australian Research Council (D.R.B. and T.P.H.), the Swedish International Development Cooperation Agency (Sida), the Swedish Research Council for the Environment, Agricultural Sciences and Spatial Planning (Formas) (C.F. and M.N.), and the Packard Foundation (Resilience Alliance).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.R.B. (david.bellwood@jcu.edu.au).

A membrane protein required for dislocation of misfolded proteins from the ER

Brendan N. Lilley & Hidde L. Ploegh

Department of Pathology, Harvard Medical School, Boston, Massachusetts, 02115, USA

After insertion into the endoplasmic reticulum (ER), proteins that fail to fold there are destroyed. Through a process termed dislocation such misfolded proteins arrive in the cytosol, where ubiquitination, deglycosylation and finally proteasomal proteolysis dispense with the unwanted polypeptides. The machinery involved in the extraction of misfolded proteins from the ER is poorly defined. The human cytomegalovirus-encoded glycoproteins US2 and US11 catalyse the dislocation of class I major histocompatibility complex (MHC) products, resulting in their rapid degradation. Here we show that US11 uses its transmembrane domain to recruit class I MHC products to a human homologue of yeast Der1p, a protein essential for the degradation of a subset of misfolded ER proteins. We show that this protein, Derlin-1, is essential for the degradation of class I MHC molecules catalysed by US11, but not by US2. We conclude that Derlin-1 is an important factor for the extraction of certain aberrantly folded proteins from the mammalian ER.

Polypeptide chains destined for the secretory pathway emerge in the lumen of the ER, where the process of folding is carefully monitored to ensure that misfits are either repaired or destroyed. How cells make the distinction between polypeptides that have yet to attain their proper conformation and those that have exhausted their folding options is not entirely clear¹, but most terminally misfolded polypeptides are transferred from the ER to the cytosol, where they are destroyed by the proteasome in a ubiquitin-dependent manner². Proteins are thought to leave the ER through the Sec61 complex^{3–7}, but the involvement of the latter has yet to be demonstrated conclusively. Substrates exposed to the cytosol are acted upon by ER-associated components of the ubiquitin conjugation machinery^{8–10}, extracted from the ER membrane by the AAA ATPase Cdc48p and its associated cofactors, Ufd1p and Npl4p (refs 11–14), and degraded by the 26S proteasome^{9,15}. Although some cytosolic components required for degradation have been characterized in mammalian cells^{2,11}, ER membrane proteins directly involved in the dislocation reaction remain to be identified by functional criteria.

Human cytomegalovirus encodes five proteins that disarm class I MHC restricted antigen presentation, thus avoiding detection by the immune system¹⁶. Two human cytomegalovirus proteins, US2 and US11, catalyse dislocation, the rapid transfer of the class I MHC heavy chain (HC) from the ER to the cytosol, where N-glycanase, ubiquitin-conjugating enzymes and finally the proteasome act on them^{3,17–20}. This pathway resembles the mechanism by which mammalian cells dispose of misfolded proteins, with US2 and US11 conferring selectivity for class I molecules and accelerating the rate constant of the dislocation reaction.

The transmembrane domain (TMD) of US11 is essential for its ability to dislocate the class I HC²¹. The presence of a single glutamine residue in the TM segment (Gln 192) suggested its involvement in interactions of US11 with host proteins in the plane of the membrane, confirmed by the inability of the Gln192Leu (US11_{Q192L}) mutant to sustain dislocation of the class I HC²¹. We exploit the availability of this US11 point mutant in an affinity purification approach aimed at isolating proteins that interact with wild-type US11 (US11_{WT}) but not with mutant US11 (US11_{Q192L}). We identify a human homologue of yeast Der1p, a protein required

for the degradation of a misfolded ER luminal protein, CPY* (ref. 22), as the partner essential for US11 to perform its function. In an accompanying paper, Ye *et al.* used a different approach and have reached a similar conclusion²³. Whereas US11_{WT} recruits the class I HC to the newly identified protein, which we call Derlin-1, US11_{Q192L} fails to do so. A dominant-negative version of Derlin-1 impedes class I HC dislocation mediated by US11 but not by US2. Combined, our results identify, by functional criteria, a mammalian integral membrane protein required for the dislocation of certain misfolded proteins from the ER to the cytosol.

Identification of US11-associated proteins

We developed an affinity purification approach that allowed us to examine cellular proteins associated with US11_{WT} and US11_{Q192L}. The purification strategy employed an amino-terminal affinity tag composed of a haemagglutinin (HA) epitope followed by a short spacer, a tobacco etch virus (TEV) protease cleavage site and an additional spacer attached to the luminal domain and remainder of US11 (Fig. 1a). HA-US11_{WT} or HA-US11_{Q192L} behaved like their untagged counterparts in their ability to cause class I HC dislocation, and migrated on SDS-polyacrylamide-gel electrophoresis (SDS-PAGE) at the expected molecular mass of 42 kDa (Fig. 1b).

We performed a large-scale purification from digitonin lysates of control U373 cells, and from cells that expressed either HA-US11_{WT} or HA-US11_{Q192L}. Both Coomassie blue staining and subsequent analysis by tandem mass spectrometry (MS/MS) showed that a significant number of polypeptides associated equally with HA-US11_{WT} and HA-US11_{Q192L} (Fig. 1c) and therefore that associations with this group of proteins are unlikely to account for the mutant phenotype of US11_{Q192L}. These proteins include ER chaperones such as immunoglobulin heavy-chain binding protein (BiP), calnexin, the AAA ATPase p97 and three subunits of the oligosaccharide transferase complex (OST48, ribophorin I and ribophorin II). The significance of these interactions will be addressed elsewhere. Consistent with our previous findings²¹ was our observation that the class I HC co-purified with HA-US11_{Q192L}. One 22-kDa polypeptide associated specifically with HA-US11_{WT} and not with HA-US11_{Q192L} (Fig. 1c). MS/MS analysis of a tryptic digest of the 22-kDa protein identified two peptides from a

predicted open reading frame, MGC3067 (NCBI GeneID 79139), which encodes a small, hydrophobic protein of 251 amino acids predicted to span the lipid bilayer four times, with both its amino and carboxy terminus in the cytosol (Fig. 1d). Immunoblots using an antiserum raised against peptides from the predicted protein sequence of MGC3067 show its presence at equivalent levels in HA-US11_{WT} and HA-US11_{Q192L} cell lines (Fig. 1e, left panel). The 22 kDa protein associates with HA-US11_{WT} but not with HA-US11_{Q192L} (Fig. 1e, right panel).

MGC3067 encodes an evolutionarily conserved protein with homologues in all eukaryotes examined. Although the functions of MGC3067 homologues in multicellular organisms are unknown, one of the yeast proteins that bears limited similarity (about 10% identity) to MGC3067 is Der1p, a known factor in the degradation of misfolded ER proteins^{22,24}. Another yeast open reading frame, YDR411c, encodes a protein more similar to MGC3067 (about 20% identity); like Der1p, YDR411cp is an ER-resident protein induced by ER stress^{25,26}.

MGC3067 and its homologues contain a domain named for yeast Der1p (Der1-like domain, Pfam accession no. PF04511; ref. 27). This domain is about 200 amino acids in length and is predominantly hydrophobic, with pockets of highly conserved and invariant

residues, many of which flank the predicted transmembrane regions (Fig. 2a). Because MGC3067 contains a Der1-like domain and shows similarity to Der1p, we name the protein product of MGC3067 Derlin-1 (Der1-like protein 1).

Mammals have two additional Der1-like domain-containing proteins homologous to Derlin-1, which we call Derlin-2 (NCBI GeneID 51009, also known as F-LANa; ref. 28) and Derlin-3 (NCBI GeneID 91319, murine orthologue known as IZP6; ref. 29). Derlin-2 and Derlin-3 are about 70% identical (see alignment in Supplementary Fig. S1) and probably originated from a gene duplication event in mammals. The evolutionary relationships between the Derlin proteins and their putative orthologues are shown in Fig. 2b.

Derlin-1 is a widely expressed protein, with strong signals obtained by immunoblotting from liver, spleen, pancreas, lung, thymus and ovary. Despite the presence of Derlin-1 sequences in brain cDNA libraries (NCBI UniGene cluster Hs.241576), we did not detect immunoreactivity in brain (Fig. 2c).

The inferred integral membrane disposition of Derlin-1 was confirmed by fractionation of U373 microsomes. Neither alkaline nor urea extraction affected the membrane association of Derlin-1 or the known ER membrane protein calnexin, but each removed the soluble ER protein PDI (protein disulphide isomerase) from the

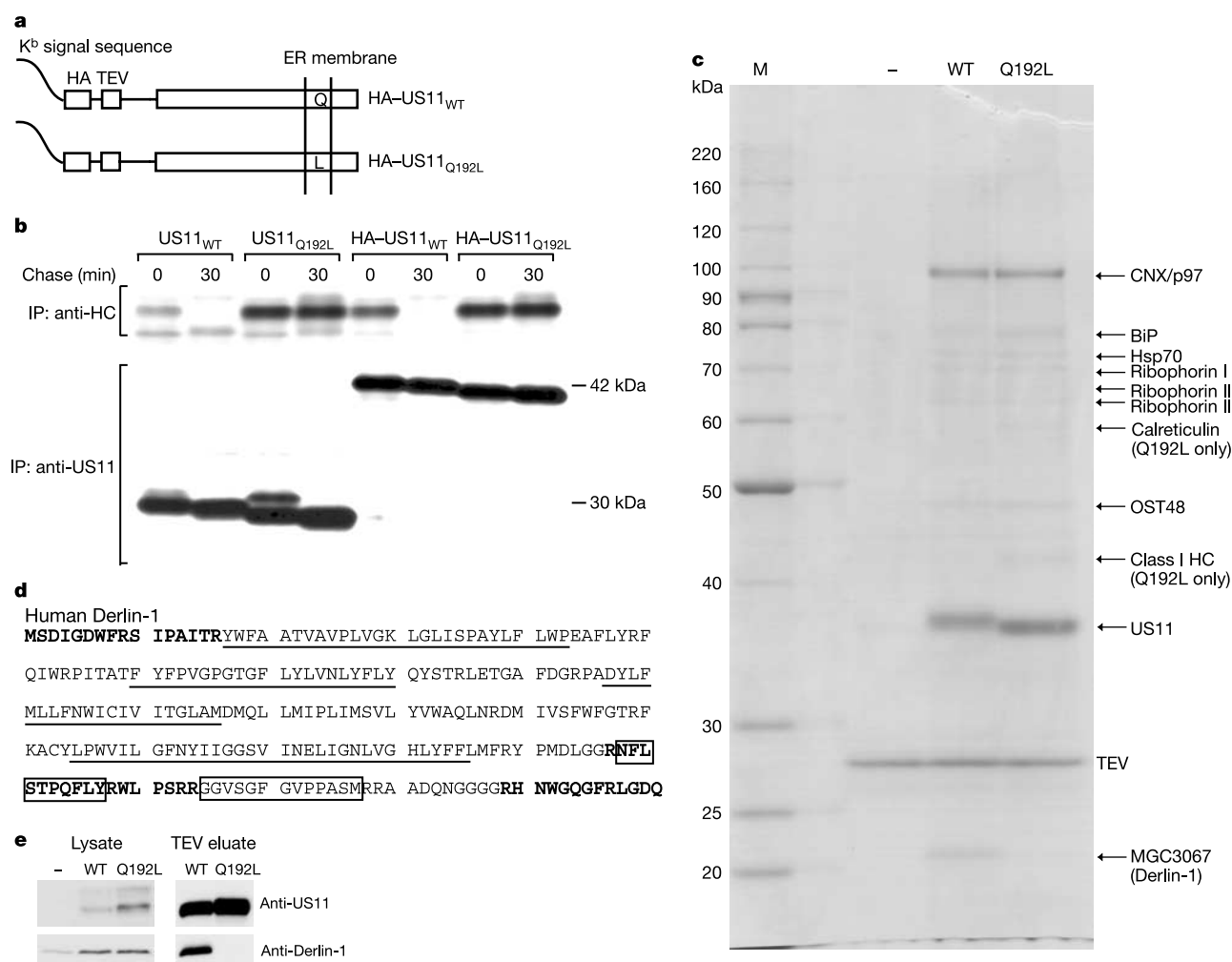


Figure 1 Identification of US11-associated proteins. **a**, Schematic representation of constructs used for affinity purification. **b**, Stability of the class I HC in cells expressing US11_{WT}, US11_{Q192L} or their epitope-tagged counterparts. IP, immunoprecipitation. **c**, Identification of HA-US11_{WT} and HA-US11_{Q192L}-associated proteins from the cell lines indicated (–, control U373 cells; WT, HA-US11_{WT}; Q192L, HA-US11_{Q192L}). Polypeptides identified by MS/MS analysis are indicated. CNX, calnexin. **d**, Sequence of human

Derlin-1. Peptides identified by MS/MS analysis are boxed, predicted TMDs (using TMpred) are underlined, and peptide sequences used for antibody production are in bold. **e**, Use of antiserum against Derlin-1 confirms the interaction of Derlin-1 with HA-US11_{WT}. Equal amounts of lysate or TEV eluate from the specified cell lines (as in **c**) were analysed by immunoblotting with anti-US11 or anti-Derlin-1 antibodies.

particulate fraction (Fig. 2d). Derlin-1 is confined to the ER, because it co-localizes with the ER chaperone calnexin (Fig. 2e). Mild proteolysis eliminated the C-terminal epitope of Derlin-1 in intact and detergent-disrupted microsomes, but the luminal protein GRP94 remained intact unless detergent was added. Intact microsomes treated with protease did not yield Derlin-1 digestion

intermediates (Fig. 2f). Thus, at least the C terminus of Derlin-1 resides in the cytosol.

Association of the class I HC with Derlin-1

The anti-Derlin-1 antiserum was used for immunoprecipitations from radiolabelled digitonin lysates of US11_{WT} and US11_{Q192L} cell

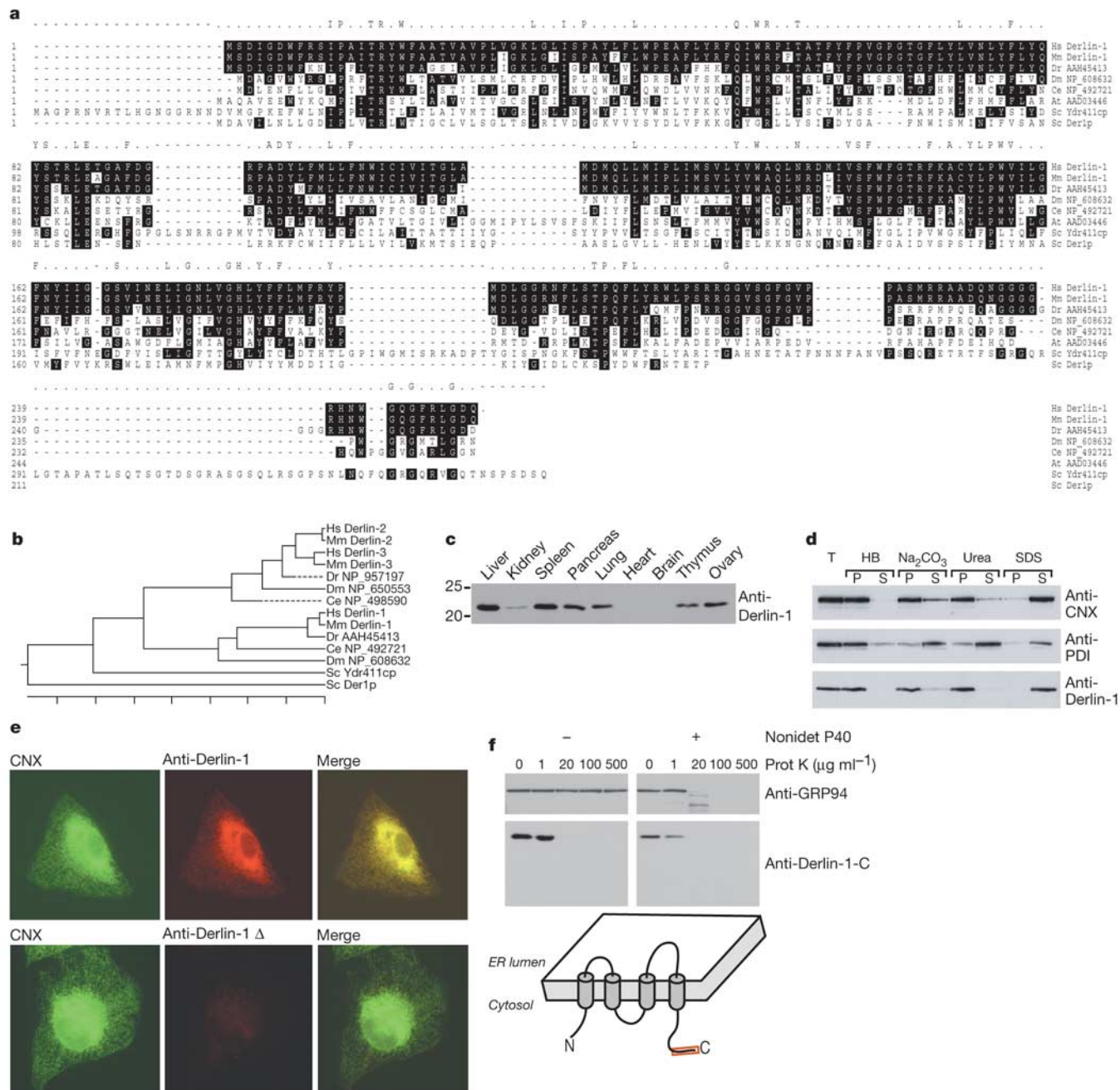


Figure 2 Characterization of Derlin-1. **a**, Alignment of Derlin-1 with putative orthologues. Sequences (GenBank accession numbers indicated) from *Danio rerio* (Dr, AAH45413), *Drosophila melanogaster* (Dm, NP_608632), *Caenorhabditis elegans* (Ce, NP_492721), *Arabidopsis thaliana* (At, AAD03446) and *Saccharomyces cerevisiae* (Sc, YDR411cp: AAB64889, Der1p: P38307) were aligned with human (Hs, NP_077271) and mouse (Mm, NP_077169) Derlin-1 by using ClustalW. Residues identical to those in human Derlin-1 are shaded, and the consensus above the aligned sequences indicates residues that are identical in six or more of the aligned sequences. **b**, Phylogenetic analysis of Der1-domain-containing proteins generated by using ClustalW and the Megalign program. The species of origin and GenBank accession numbers are shown or are as follows: Hs Derlin-2, NP_057125; Mm Derlin-2, NP_291040; Hs Derlin-3, AAH57830;

Mm Derlin-3, BAB32788. **c**, Analysis of Derlin-1 expression in mouse tissues by immunoblotting. **d**, Derlin-1 is an integral membrane protein. Immunoblot analysis of microsomes treated with the indicated reagents (HB, homogenization buffer) performed on total (T), particulate (P) and soluble (S) fractions by using antibodies against calnexin (anti-CN), PDI and Derlin-1. **e**, Derlin-1 is an ER-resident protein. U373 cells were stained with antibodies against calnexin (CN), Derlin-1 (anti-Derlin-1) or Derlin-1 antiserum depleted of specific antibodies (anti-Derlin-1 Δ). Merging the images shows the co-localization of Derlin-1 with calnexin. **f**, Analysis of Derlin-1 topology by protease protection and immunoblotting with anti-GRP94 and anti-Derlin-1-C antibodies. The removal of the C terminus of Derlin-1 in the absence of detergent (Nonidet P40) is consistent with the topology proposed.

lines. Labelled US11_{WT} immunoprecipitates with Derlin-1 but US11_{Q192L} does not (Fig. 3b, compare lanes 1 and 2 with 5 and 6), which is consistent with the results of the large-scale purification. Re-immunoprecipitation from the material captured by the anti-Derlin-1 antibodies with anti-US11 serum confirms the presence of US11_{WT} and the absence of US11_{Q192L} (see Fig. 5e, lower panel). Labelled US11_{WT} associates with Derlin-1 very soon after synthesis (Fig. 3b), indicating that this complex forms rapidly. The detection of radiolabelled Derlin-1 requires prolonged exposure of the autoradiogram (data not shown), possibly reflecting a low level of Derlin-1 synthesis relative to US11_{WT} and the class I HC during the 30 min of pulse labelling.

Does the class I HC substrate also occur in a complex that contains Derlin-1? We first showed in a pulse-chase experiment that the class I HC disappears rapidly from US11_{WT} cells (Fig. 3a,

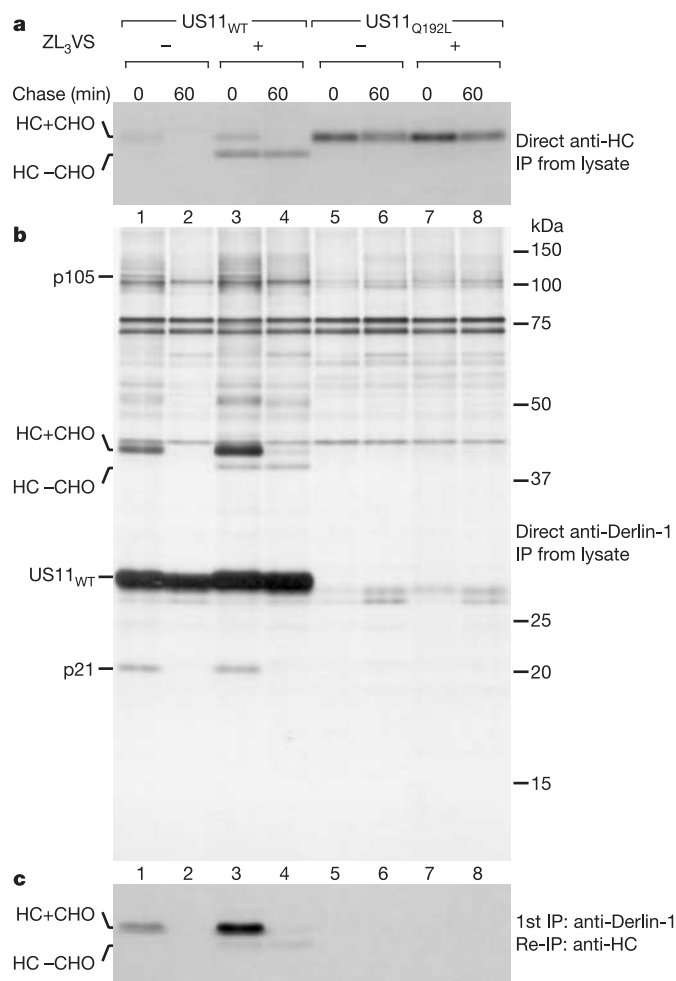


Figure 3 Association of US11_{WT} and the class I HC with Derlin-1. **a**, Digitonin lysates from the indicated cell lines were used for direct immunoprecipitation (IP) with anti-heavy chain (anti-HC) antibodies. Treatment with proteasome inhibitors (ZL₃VS) leads to the accumulation of deglycosylated class I HC (HC — CHO) only in US11_{WT} cells. The majority of the class I HC is in the deglycosylated form after the pulse. **b**, Digitonin lysates from the indicated cell lines were immunoprecipitated with anti-Derlin-1 antibodies. Half of the recovered material was analysed directly and the remaining half was used for re-immunoprecipitation with anti-heavy chain antibodies (shown in **c**). Only US11_{WT} is recovered with Derlin-1, and predominantly glycosylated (HC + CHO) class I HC is recovered with Derlin-1 only in US11_{WT} cells. Additional radiolabelled polypeptides of 105 kDa (p105) and 21 kDa (p21) associate with Derlin-1 in US11_{WT} cells but not in US11_{Q192L} cells. **c**, Re-immunoprecipitation of the class I HC (anti-HC) from Derlin-1 immunoprecipitates in **b** confirms the presence of the class I HC. The autoradiogram in **a** was exposed for 1 day; those in **b** and **c** were exposed for 6 days.

lanes 1 and 2). Inclusion of proteasome inhibitors in pulse-chase experiments with US11 cells causes the transient accumulation of a deglycosylated class I HC intermediate whose N-linked glycan has been removed by a cytoplasmic N-glycanase (Fig. 3a, lanes 3 and 4; ref. 17). The US11_{Q192L} mutant is incapable of accelerating degradation of the class I HC (Fig. 3a, lanes 5–8). In cells expressing US11_{WT}, immunoprecipitations for Derlin-1 also recovered the class I HC (Fig. 3b, lane 1). The class I HC associates with Derlin-1 at early times after synthesis, but later, coincident with the degradation of the class I HC, this association is lost (Fig. 3b, lanes 1 and 2). In cells that express US11_{Q192L}, no class I HC is associated with Derlin-1 (Fig. 3b, lanes 5 and 6). Immunoprecipitations with either preimmune serum or immune serum that had been depleted of specific anti-Derlin-1 antibodies did not recover any detectable amount of class I HC or US11_{WT} (data not shown). US11 therefore uses its TMD to recruit the class I HC into a complex with Derlin-1.

In US11_{WT} cells treated with proteasome inhibitor, we find increased amounts of glycosylated class I HC in association with Derlin-1 relative to non-treated cells (Fig. 3b, compare lanes 3 and 1). Indeed, prolonged treatment with proteasome inhibitor delays the dislocation of class I HC to the cytosol³⁰. When Derlin-1 was

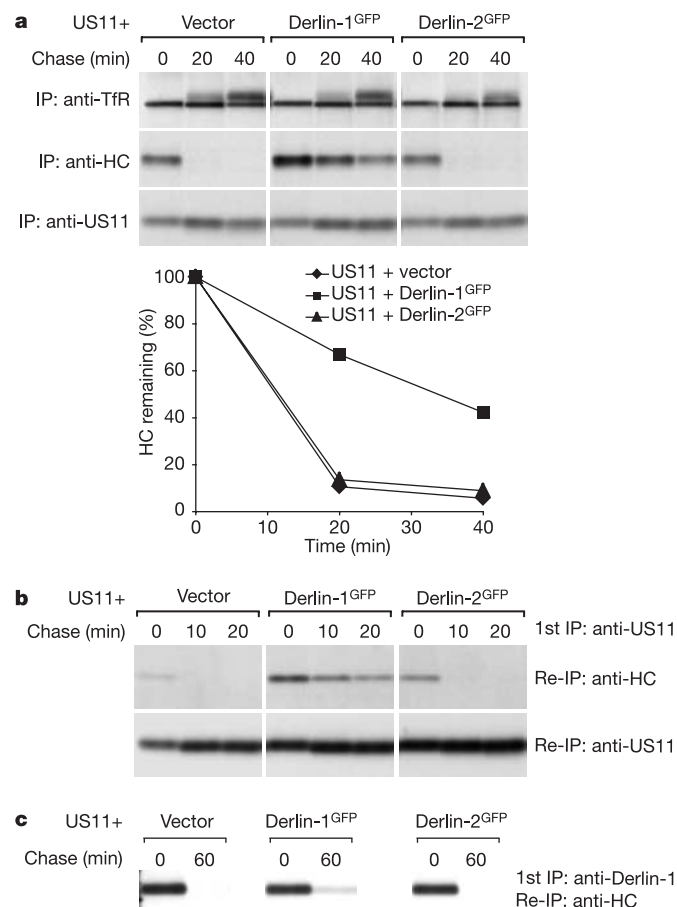


Figure 4 Expression of a Derlin-1 dominant-negative construct impedes US11-mediated class I HC degradation. **a**, The stability of the class I HC in US11_{WT} cells is substantially increased by the expression of Derlin-1^{GFP}. The graph shows quantification of the experiment. Immunoprecipitation (IP) for transferrin receptor (anti-TfR) shows that glycoprotein maturation is not affected in cells that express Derlin-1^{GFP} or Derlin-2^{GFP}. **b**, The complex of US11_{WT} and the class I HC persists when Derlin-1^{GFP} is expressed. Digitonin lysates were immunoprecipitated for US11, followed by re-immunoprecipitation with anti-HC and anti-US11 antibodies. **c**, The complex of the class I HC and Derlin-1 persists in the presence of Derlin-1^{GFP}. Digitonin lysates were immunoprecipitated for Derlin-1, followed by re-immunoprecipitation with anti-HC antibodies.

immunoprecipitated from cells treated with proteasome inhibitor, predominantly glycosylated class I HC was immunoprecipitated with it after the onset of the chase, despite the fact that more than 50% of the total class I HC population was in the deglycosylated form (Fig. 3, lane 3, compare panels a and b). A small amount of deglycosylated class I HC was also immunoprecipitated with Derlin-1 (Fig. 3b, c, lanes 3 and 4). Therefore, although Derlin-1 associates preferentially with glycosylated class I HC, it can interact with class I HC that has been exposed to the cytosol.

In cells that express US11_{WT}, several other proteins occur in association with Derlin-1 in addition to US11_{WT} and the class I HC. Proteins of about 105 and 21 kDa associated with Derlin-1 after the onset of the chase in cells expressing US11_{WT} but not in those expressing US11_{Q192L} (Fig. 3b, compare lanes 1–4 with lanes 5–8). The 105-kDa species is not the AAA ATPase, p97 (ref. 11) as demonstrated by re-immunoprecipitation (data not shown). These labelled species were not recovered with Derlin-1 at later time points. Although we do not yet know the identity of these proteins, their association with Derlin-1 in US11_{WT} cells indicates that they, too, might have a function in dislocation of the class I HC mediated by US11.

Derlin-1 is required for dislocation

To test the function of Derlin-1 in the US11-mediated dislocation of class I HC, we created a dominant-negative construct of Derlin-1 and, as a specificity control, Derlin-2. On the basis of the structure of Derlin-1 and related proteins, we proposed that the addition of a folded domain to what is predicted to be a flexible cytoplasmic tail would interfere with interactions on the cytoplasmic side of the ER but would leave intact the intramembrane interaction with the US11 TMD. We therefore added green fluorescent protein (GFP) to the C terminus of Derlin-1 and Derlin-2 (Derlin-1^{GFP} and Derlin-2^{GFP}) and expressed these constructs in US11_{WT} cells. Whereas Derlin-1^{GFP} was expressed at a level roughly equivalent to that of endogenous Derlin-1, Derlin-2^{GFP} was significantly overexpressed relative to endogenous Derlin-2, as assessed by immunoblotting

(Supplementary Fig. S2a). Derlin-1 and Derlin-2^{GFP} localize to the ER (Supplementary Fig. S2b), and their expression did not obviously perturb ER structure³¹ (data not shown) or trafficking of the transferrin receptor as assessed by N-glycan maturation (Fig. 4a, top panels).

In US11_{WT} cells expressing Derlin-1^{GFP}, the class I HC was more stable than in US11_{WT} cells expressing Derlin-2^{GFP} (Fig. 4a, middle panels). The half-life of the class I HC in US11_{WT} cells is extended from less than 5 min (ref. 17) to more than 30 min in the presence of Derlin-1^{GFP} (Fig. 4a).

Shortly after synthesis, but before dislocation, the class I HC exists as part of a complex in the ER that includes US11_{WT} (ref. 21) and Derlin-1 (Fig. 3). As dislocation proceeds, the class I HC is lost from this complex. In cells that express Derlin-1^{GFP} we observe a prolonged association of the class I HC with both US11_{WT} (Fig. 4b) and with Derlin-1 (Fig. 4c). The presence of Derlin-1^{GFP} thus delays the exposure of the class I HC to the cytosol. Expression of Derlin-1^{GFP} does not preclude binding of the class I HC substrate to either US11_{WT} or Derlin-1 itself.

Specificity of Derlin-1 action

Both US11 and US2 catalyse dislocation of the class I HC, resulting in intermediates and a final outcome that are similar^{3,17}. Having established that Derlin-1^{GFP} blocks dislocation of class I HC molecules in US11_{WT} cells, we examined its effects on US2-dependent dislocation of class I molecules.

At levels of Derlin-1^{GFP} expression that block US11-dependent dislocation (Supplementary Fig. S2a, c), the class I HC molecules in cells that express US2 (US2 cells) are not affected (Fig. 5a, middle panels). Thus, the inhibition of dislocation by Derlin-1^{GFP} in US11_{WT} cells is specific and cannot be attributed to a systemic inhibitory effect on ER function. In US2 cells, we do not observe an association of the class I HC with Derlin-1 (Fig. 5e), which is consistent with a lack of Derlin-1 involvement in US2-mediated class I HC dislocation. When US2 cells are treated with proteasome inhibitors, the deglycosylated class I HC is observed, as it is in

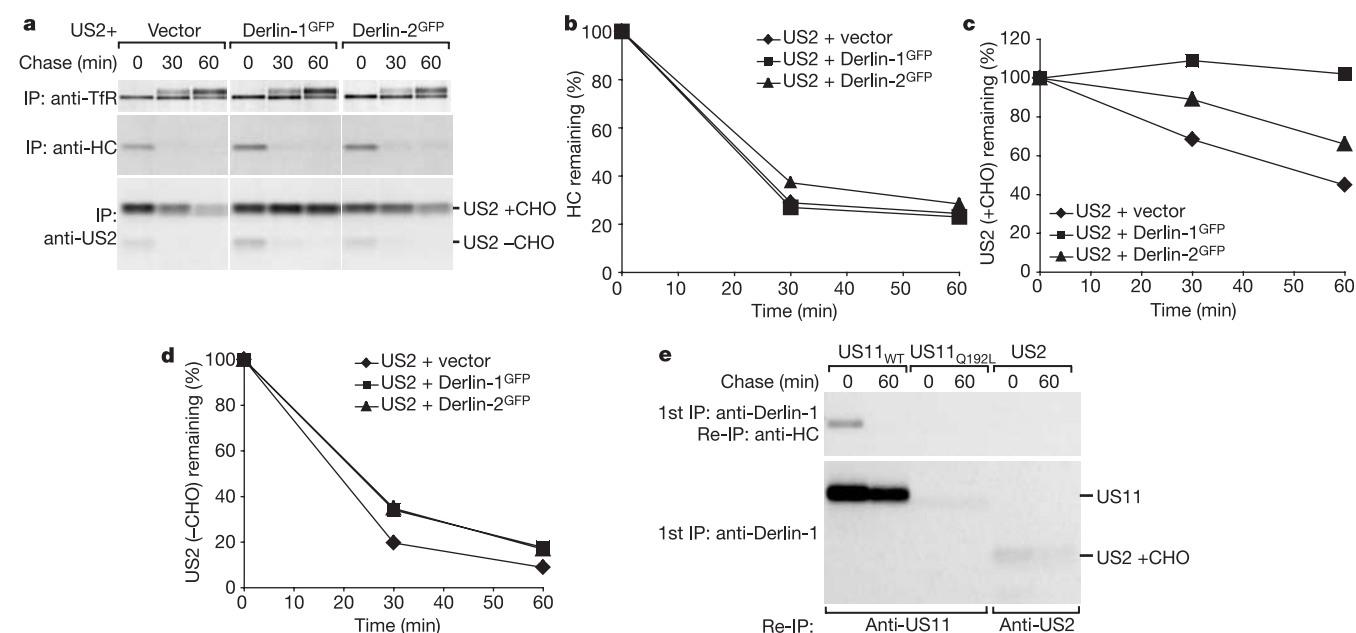


Figure 5 Expression of a Derlin-1 dominant-negative construct does not inhibit US2-mediated class I HC degradation, but blocks degradation of the US2 protein itself. **a**, Stability of the class I HC is not affected in US2 cells expressing Derlin-1^{GFP} compared with control US2 cells, but degradation of glycosylated US2 protein (US2+CHO) is blocked in the presence of Derlin-1^{GFP}. Degradation of the non-glycosylated US2 (US2-CHO) is not affected. Immunoprecipitation (IP) for transferrin receptor (anti-TfR) shows that

glycoprotein maturation is not affected in cells that express Derlin-1^{GFP} or Derlin-2^{GFP}. **b–d**, Quantification of the results in **a**: **b**, class I HC; **c**, US2+CHO; **d**, US2-CHO. **e**, The US2 protein, but not the class I HC, associates with Derlin-1 in US2 cells. Digitonin lysates from the indicated cell lines were immunoprecipitated with anti-Derlin-1 antibodies followed by re-immunoprecipitation first with anti-HC (upper panel) antibodies followed by anti-US11 or anti-US2 antibodies (lower panel).

US11_{WT} cells³. However, in US2 cells treated with proteasome inhibitors, we did not observe deglycosylated class I HC in association with Derlin-1 (data not shown), whereas we did observe a small amount of this class I HC intermediate associated with Derlin-1 in US11_{WT} cells (Fig. 3b, lanes 3 and 4). This supports the idea that, in US11_{WT} cells, the interaction between Derlin-1 and the deglycosylated class I HC is specific and not due to an interaction after lysis. The differential inhibition of US11- versus US2-mediated dislocation by Derlin-1^{GFP} suggests that US2 and US11 exploit different pathways³² for dislocation of the same substrate proteins.

Analysis of US2 cells also allowed us to address the fate of US2 itself, a short-lived type I membrane protein that lacks a cleavable signal sequence³³. The US2 protein exists in two forms in the cell: an ER-inserted, glycosylated species, and a non-glycosylated, cytosolic version. The cytosolic, non-glycosylated form is derived from the failure of a fraction of the newly synthesized US2 to insert into the ER. The ER-resident form of US2 has a short half-life, and represents an additional ER degradation substrate³³.

In US2 cells that express Derlin-1^{GFP}, US2 is stabilized but the degradation of class I HC molecules and non-glycosylated US2 continues (Fig. 5a–d). A small amount of glycosylated US2 associates with Derlin-1 (Fig. 5e). The degradation of two type I membrane proteins, the class I HC in US11_{WT} cells and the US2 protein, therefore requires Derlin-1. We have not yet identified a functional role for Derlin-2 in the degradation of ER proteins, but the Derlin-2^{GFP} fusion protein in no way impairs dislocation of the class I HC in either US11_{WT} cells (Fig. 4) or US2 cells (Fig. 5).

Discussion

We identified Derlin-1 as an ER membrane protein essential for US11-mediated dislocation of the class I HC. Our approach rests on the assumption that the active form of US11 should interact with partner proteins no longer available to the inactive US11_{Q192L}. Indeed, affinity purification identified Derlin-1 as a protein that meets this criterion. The use of a dominant-negative version of Derlin-1 shows its direct involvement in US11-mediated dislocation. We emphasize the importance of a functional readout of US11 activity, as exemplified both by the distinction between US11_{WT} and US11_{Q192L} and by the ability of a dominant-negative version of Derlin-1 to block US11-dependent dislocation. Although mere interactions between proteins involved in dislocation might suggest functional relevance, as observed for US11_{WT} class I HC and Derlin-1, these observations are confirmed by analysis of US11_{Q192L} and the effects of the dominant-negative form of Derlin-1 on class I HC dislocation.

Derlin-1 is weakly similar to yeast Der1p, whose mechanistic contribution to ER degradation is unknown. The degradation of only a subset of ER substrates requires the involvement of Der1p, which is consistent with the notion that multiple pathways operate to clear misfolded proteins from the ER^{22,24,34–38}. We show that Derlin-1 mediates the degradation of two different type I membrane proteins, the class I HC and US2. At present we know of no additional proteins that require Derlin-1 for their degradation, but we consider it likely that Derlin-1 acts to degrade cellular ER proteins independently of the US11 and US2 systems studied here. The possible role(s) of additional Der1-like domain-containing proteins in ER degradation have yet to be examined. These proteins might serve a function similar to that of Der1p and Derlin-1 but might operate on distinct groups of substrate proteins. In view of the diverse nature of proteins that exist in the secretory pathway, there might be factors that operate independently of Derlin-1 to achieve quality control.

The Sec61 complex might be the conduit by which proteins exit the ER^{3–7}. However, the structure of the analogous SecY complex from archaea suggests a strict upper limit to the diameter of the pore that spans the lipid bilayer³⁹. What is currently known about the dislocation reaction is hard to reconcile with the observed

dimensions of the SecY complex. Class I HC fusion proteins prepared with the tightly folded domain of GFP, or with dihydrofolate reductase stabilized by methotrexate analogues, are still dislocated with kinetics seen for the unmodified class I HC^{40,41}, as is the high-mannose-bearing class I HC²⁰, further indicating a dislocation pore of considerable size.

We observe a significant amount of the glycosylated class I HC in association with Derlin-1 before dislocation to the cytosol. When cells are treated with proteasome inhibitor, some deglycosylated class I HC occurs in association with Derlin-1. Expression of a Derlin-1 dominant-negative construct causes glycosylated class I HC to persist in a complex with US11_{WT} and Derlin-1, which is indicative of class I HC retention in the ER. Derlin-1 is therefore positioned in the ER membrane in such a way that it can interact with substrate proteins both before and immediately after their dislocation to the cytosol. The presence in Derlin-1 of four transmembrane segments now suggests further possibilities for the creation of a protein-conducting channel—the dislocon—that might be gated through the oligomerization of Derlin-1 or through the association of Derlin-1 with accessories yet to be identified. The formation of this channel might indeed be independent of Sec61.

US11 is a small 215-residue glycoprotein that consists of at least two modules. The luminal domain is sufficient for interaction with the class I HC²¹, whereas the TMD is required for interaction with Derlin-1. Even though the interaction between the US11 TMD and Derlin-1 might be indirect, contacts formed by Gln 192 in US11 are essential for the interaction between US11 and Derlin-1. The US11 protein is remarkable for the speed with which it targets the class I HC molecules for degradation: within minutes of its synthesis, the class I HC is destroyed. We propose that US11 does so by recruitment of the class I HC to Derlin-1. For other ER degradation substrates, they must first be recognized as misfolded or unassembled, and then must be targeted to the machinery that performs the dislocation reaction. We propose that the recruitment of such misfolded proteins to Der1-like domain-containing proteins represents a scheme for clearing the ER of the requisite range of substrate proteins. □

Methods

DNA constructs and cell lines

HA-US11_{WT} and HA-US11_{Q192L} were constructed with standard methods, and sequence-verified constructs were transferred into a pLNCX-based vector²¹ for retrovirus production. The murine K^b signal sequence was included to direct the proteins to the ER. Human Derlin-1 and Derlin-2 were cloned from U373 astrocytoma cDNA by using polymerase chain reaction with reverse transcription, and GFP fusions were made by the insertion of Derlin-1 and Derlin-2 complementary DNAs into the pEGFP-N1 vector (Clontech). The Derlin-1^{GFP} and Derlin-2^{GFP} constructs were subcloned into the pLHCX retroviral expression vector (Clontech). The US11_{WT} and US11_{Q192L} cell lines have been described^{21,42}. US2 cells were derived by transducing U373 cells with a retroviral construct containing the full-length US2 gene, followed by antibiotic selection and cloning by limiting dilution. Methods for the production of retroviruses, infection of U373 cells and subsequent antibiotic selection have been reported²¹. All U373 astrocytoma cell lines used were maintained in culture as described⁴². Cells transduced with LHCX-based vectors were selected with 125 µg ml^{−1} hygromycin B (Roche).

Antibodies and immunoblotting

Most of the antibodies used in this study have been reported^{21,43}. Antibodies against transferrin receptor (H68.4) and p97 were purchased from Zymed and Research Diagnostics, respectively. Alexa-488-conjugated goat anti-mouse and Alexa-568-conjugated goat anti-rabbit were from Molecular Probes. Anti-Derlin-1 and anti-Derlin-2 antisera were generated by immunizing rabbits with peptides coupled to keyhole-limpet haemocyanin through an added Cys residue. The Derlin-1 sequences used were SDIGDWFRSIPATR(C), (C)RNFLSTPQFLYRWLPSRR and (C)RHNWGGQFRLGDQ. Derlin-2 sequences used were AYQSLRLEYLQIPVSR(C), (C)KAIFDTPDEDPNYN and (C)EERPGGFAWGEQRLGG. Antibodies specific to the C terminus of Derlin-1 (RHNWGGQFRLGDQ) were affinity purified as described⁴⁴. Affinity-purified anti-Derlin-1-C antibodies were used for immunoprecipitations, examination of Derlin-1 topology, and immunohistochemical analysis. Crude anti-Derlin-1 serum was used for immunoblotting. Anti-GFP and anti-GRP94 sera were generated by immunizing rabbits with bacterially expressed GFP and GRP94, respectively. Immunoblotting experiments were performed as described²¹.

Purification of HA-US11_{WT} and HA-US11_{Q192L}, and MS/MS analysis

Cells (10⁸) were lysed in 30 ml of ice-cold lysis buffer (1% digitonin in 25 mM Tris-HCl pH 7.4, 150 mM NaCl, 5 mM MgCl₂, 1 mM CaCl₂ with 1 mM phenylmethylsulphonyl fluoride and complete protease inhibitors (Roche)). The lysate was cleared of debris by centrifugation at 20,000 g for 15 min and was added to 0.35 ml of 12CA5-coupled Sepharose²¹. The affinity resin was washed with 35 ml of wash buffer (composition as lysis buffer, except with 0.1% digitonin and without protease inhibitors). Bound material was eluted by treatment of the resin for 1 h at 25 °C with 80 units of TEV protease (Invitrogen) in 0.25 ml of wash buffer. The eluate was exchanged into 20 mM NH₄CO₃ pH 8.0 with 0.1% SDS with the use of Sephadex G-25 resin and was then freeze-dried. Polypeptides were separated by 10% reducing SDS-PAGE and were revealed by Coomassie blue staining. Polypeptides of interest were excised and subjected to trypsinolysis as described⁴⁵. Digested samples were separated by liquid chromatography and analysed by MS/MS as described⁴⁶. After data acquisition and processing, MS/MS data were searched against the NCBI non-redundant database using the MASCOT program (Matrix Science).

Pulse-chase experiments and SDS-PAGE

Methods for the treatment of cells with the proteasome inhibitor ZL₃VS (ref. 19), pulse labelling, cell lysis, immunoprecipitation and analysis of class I HC stability in US11 and US2 cells have been described²¹. Preparation of digitonin lysates and re-immunoprecipitation of specified polypeptides from material captured from digitonin lysates was performed as described²¹. N-Ethylmaleimide was included during digitonin lysate preparation at a concentration of 2.5 mM. Methods for SDS-PAGE and fluorography have been described⁴⁷. In experiments examining proteins associated with Derlin-1, the pulse labelling period was extended to 30 min, and the cells were chased for 60 min.

Analysis of Derlin-1

Microsomes from U373 cells were produced as described³² and were treated with 0.1 M Na₂CO₃, pH 11.6, 2.5 M urea or 1% SDS as reported²⁴. Treatment of U373 microsomes with protease K was performed as described⁴⁸. For the analysis of Derlin-1 tissue distribution, organs from a PBS-perfused C57/BL6 mouse were removed and homogenized in 100 mM Tris-HCl pH 7.6 with 8 M urea and 1% SDS. Equal amounts of protein were precipitated with trichloroacetic acid and a 25 µg sample from each tissue source was analysed by reducing 12% SDS-PAGE followed by immunoblotting. Another gel was run in parallel and was analysed by Coomassie blue staining to ensure equivalent protein loading. Immunohistochemical analysis of Derlin-1 and calnexin localization in U373 cells was performed with established methods²¹. Anti-Derlin-1 serum that had been depleted of specific antibodies by affinity purification was used as a control.

Received 8 March; accepted 21 April 2004; doi:10.1038/nature02592.

- Ellgaard, L. & Helenius, A. Quality control in the endoplasmic reticulum. *Nature Rev. Mol. Cell Biol.* **4**, 181–191 (2003).
- Kostova, Z. & Wolf, D. H. For whom the bell tolls: protein quality control of the endoplasmic reticulum and the ubiquitin-proteasome connection. *EMBO J.* **22**, 2309–2317 (2003).
- Wiertz, E. J. *et al.* Sec61-mediated transfer of a membrane protein from the endoplasmic reticulum to the proteasome for destruction. *Nature* **384**, 432–438 (1996).
- Bebok, Z., Mazzochi, C., King, S. A., Hong, J. S. & Sorscher, E. J. The mechanism underlying cystic fibrosis transmembrane conductance regulator transport from the endoplasmic reticulum to the proteasome includes Sec61beta and a cytosolic, deglycosylated intermediary. *J. Biol. Chem.* **273**, 29873–29878 (1998).
- de Virgilio, M., Weninger, H. & Ivessa, N. E. Ubiquitination is required for the retro-translocation of a short-lived luminal endoplasmic reticulum glycoprotein to the cytosol for degradation by the proteasome. *J. Biol. Chem.* **273**, 9734–9743 (1998).
- Plempner, R. K., Bohmler, S., Bordallo, J., Sommer, T. & Wolf, D. H. Mutant analysis links the translocon and BiP to retrograde protein transport for ER degradation. *Nature* **388**, 891–895 (1997).
- Plempner, R. K. *et al.* Genetic interactions of Hrd3p and Der3p/Hrd1p with Sec61p suggest a retro-translocation complex mediating protein transport for ER degradation. *J. Cell Sci.* **112**, 4123–4134 (1999).
- Sommer, T. & Jentsch, S. A protein translocation defect linked to ubiquitin conjugation at the endoplasmic reticulum. *Nature* **365**, 176–179 (1993).
- Hiller, M. M., Finger, A., Schweiger, M. & Wolf, D. H. ER degradation of a misfolded luminal protein by the cytosolic ubiquitin-proteasome pathway. *Science* **273**, 1725–1728 (1996).
- Bays, N. W., Gardner, R. G., Seelig, L. P., Joazeiro, C. A. & Hampton, R. Y. Hrd1p/Der3p is a membrane-anchored ubiquitin ligase required for ER-associated degradation. *Nature Cell Biol.* **3**, 24–29 (2001).
- Ye, Y., Meyer, H. H. & Rapoport, T. A. The AAA ATPase Cdc48/p97 and its partners transport proteins from the ER into the cytosol. *Nature* **414**, 652–656 (2001).
- Hitchcock, A. L. *et al.* The conserved npl4 protein complex mediates proteasome-dependent membrane-bound transcription factor activation. *Mol. Biol. Cell* **12**, 3226–3241 (2001).
- Jarosh, E. *et al.* Protein dislocation from the ER requires polyubiquitination and the AAA-ATPase Cdc48. *Nature Cell Biol.* **4**, 134–139 (2002).
- Bays, N. W., Wilhovsky, S. K., Goradia, A., Hodgkiss-Harlow, K. & Hampton, R. Y. HRD4/NPL4 is required for the proteasomal processing of ubiquitinated ER proteins. *Mol. Biol. Cell* **12**, 4114–4128 (2001).
- Hampton, R. Y., Gardner, R. G. & Rine, J. Role of 26S proteasome and HRD genes in the degradation of 3-hydroxy-3-methylglutaryl-CoA reductase, an integral endoplasmic reticulum membrane protein. *Mol. Biol. Cell* **7**, 2029–2044 (1996).
- Tortorella, D., Gewurz, B. E., Furman, M. H., Schust, D. J. & Ploegh, H. L. Viral subversion of the immune system. *Annu. Rev. Immunol.* **18**, 861–926 (2000).
- Wiertz, E. J. *et al.* The human cytomegalovirus US11 gene product dislocates MHC class I heavy chains from the endoplasmic reticulum to the cytosol. *Cell* **84**, 769–779 (1996).

- Shamu, C. E., Flierman, D., Ploegh, H. L., Rapoport, T. A. & Chau, V. Polyubiquitination is required for US11-dependent movement of MHC class I heavy chain from endoplasmic reticulum into cytosol. *Mol. Biol. Cell* **12**, 2546–2555 (2001).
- Shamu, C. E., Story, C. M., Rapoport, T. A. & Ploegh, H. L. The pathway of US11-dependent degradation of MHC class I heavy chains involves a ubiquitin-conjugated intermediate. *J. Cell Biol.* **147**, 45–58 (1999).
- Blom, D., Hirsch, C., Stern, P., Tortorella, D. & Ploegh, H. L. A glycosylated type I membrane protein becomes cytosolic when peptide:N-glycanase is compromised. *EMBO J.* **23**, 650–658 (2004).
- Lilley, B. N., Tortorella, D. & Ploegh, H. L. Dislocation of a type I membrane protein requires interactions between membrane-spanning segments within the lipid bilayer. *Mol. Biol. Cell* **14**, 3690–3698 (2003).
- Knop, M., Finger, A., Braun, T., Hellmuth, K. & Der Wolf, D. H. Der1, a novel protein specifically required for endoplasmic reticulum degradation in yeast. *EMBO J.* **15**, 753–763 (1996).
- Ye, Y. *et al.* A membrane protein complex mediates retro-translocation from the ER lumen into the cytosol. *Nature* (this issue).
- Taxis, C. *et al.* Use of modular substrates demonstrates mechanistic diversity and reveals differences in chaperone requirement of ERAD. *J. Biol. Chem.* **278**, 35903–35913 (2003).
- Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
- Travers, K. J. *et al.* Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation. *Cell* **101**, 249–258 (2000).
- Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32** (Database issue), D138–D141 (2004).
- Ying, H., Yu, Y. & Xu, Y. Cloning and characterization of F-LANa, upregulated in human liver cancer. *Biochem. Biophys. Res. Commun.* **286**, 394–400 (2001).
- Tsakahara, M., Ji, Z. S., Noguchi, S. & Tsunoo, H. A novel putative transmembrane protein, IZP6, is expressed in neural cells during embryogenesis. *Dev. Growth Differ.* **43**, 285–293 (2001).
- Story, C. M., Furman, M. H. & Ploegh, H. L. The cytosolic tail of class I MHC heavy chain is required for its dislocation by the human cytomegalovirus US2 and US11 gene products. *Proc. Natl Acad. Sci. USA* **96**, 8516–8521 (1999).
- Snapp, E. L. *et al.* Formation of stacked ER cisternae by low affinity protein interactions. *J. Cell Biol.* **163**, 257–269 (2003).
- Furman, M. H., Ploegh, H. L. & Tortorella, D. Membrane-specific, host-derived factors are required for US2- and US11-mediated degradation of major histocompatibility complex class I molecules. *J. Biol. Chem.* **277**, 3258–3267 (2002).
- Gewurz, B. E., Ploegh, H. L. & Tortorella, D. US2, a human cytomegalovirus-encoded type I membrane protein, contains a non-cleavable amino-terminal signal peptide. *J. Biol. Chem.* **277**, 11306–11313 (2002).
- Vashist, S. *et al.* Distinct retrieval and retention mechanisms are required for the quality control of endoplasmic reticulum protein folding. *J. Cell Biol.* **155**, 355–368 (2001).
- Plempner, R. K., Egner, R., Kuchler, K. & Wolf, D. H. Endoplasmic reticulum degradation of a mutated ATP-binding cassette transporter Pdr5 proceeds in a concerted action of Sec61 and the proteasome. *J. Biol. Chem.* **273**, 32848–32856 (1998).
- Walter, J., Urban, J., Volkwein, C. & Sommer, T. Sec61p-independent degradation of the tail-anchored ER membrane protein Ubc6p. *EMBO J.* **20**, 3124–3131 (2001).
- Hill, K. & Cooper, A. A. Degradation of unassembled Vph1p reveals novel aspects of the yeast ER quality control system. *EMBO J.* **19**, 550–561 (2000).
- Vashist, S. & Ng, D. T. W. Misfolded proteins are sorted by a sequential checkpoint mechanism of ER quality control. *J. Cell Biol.* **165**, 41–52 (2004).
- Van den Berg, B. *et al.* X-ray structure of a protein-conducting channel. *Nature* **427**, 36–44 (2004).
- Fiebigler, E., Story, C., Ploegh, H. L. & Tortorella, D. Visualization of the ER-to-cytosol dislocation reaction of a type I membrane protein. *EMBO J.* **21**, 1041–1053 (2002).
- Tirosh, B., Furman, M. H., Tortorella, D. & Ploegh, H. L. Protein unfolding is not a prerequisite for endoplasmic reticulum-to-cytosol dislocation. *J. Biol. Chem.* **278**, 6664–6672 (2003).
- Rehm, A., Stern, P., Ploegh, H. L. & Tortorella, D. Signal peptide cleavage of a type I membrane protein, HCMV US11, is dependent on its membrane anchor. *EMBO J.* **20**, 1573–1582 (2001).
- Tortorella, D. *et al.* Dislocation of type I membrane proteins from the ER to the cytosol is sensitive to changes in redox potential. *J. Cell Biol.* **142**, 365–376 (1998).
- Sawin, K. E., Mitchison, T. J. & Wordeman, L. G. Evidence for kinesin-related proteins in the mitotic apparatus using peptide antibodies. *J. Cell Sci.* **101**, 303–313 (1992).
- Kinter, M. & Sherman, N. E. *Protein Sequencing and Identification using Tandem Mass Spectrometry* (Wiley, New York, 2000).
- Borodovsky, A. *et al.* Chemistry-based functional proteomics reveals novel members of the deubiquitinating enzyme family. *Chem. Biol.* **9**, 1149–1159 (2002).
- Ploegh, H. L. in *Current Protocols in Protein Science* (eds Coligan, J. E., Dunn, B. M., Ploegh, H. L., Speicher, D. W. & Wingfield, P. T.) 10.2.1–10.2.8 (Wiley, New York, 1995).
- Green, N., Fang, H., Kalies, K.-U. & Canfield, V. in *Current Protocols in Cell Biology* (eds Bonifacio, J. S., Dasso, M., Harford, J. B., Lippincott-Schwartz, J. & Yamada, K. M.) 5.2.1–5.2.27 (Wiley, New York, 1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank B. Kessler and E. Spooner for the preparation of samples and assistance in analysis of mass spectrometry data; R. Tirabassi and K. Ryan for the production of anti-GFP and anti-GRP94 antisera; D. Tortorella for cell lines; and members of the Ploegh laboratory for helpful comments on the manuscript. B.N.L. is a Howard Hughes Medical Institute Predoctoral Fellow. This work was supported by the National Institutes of Health.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.L.P. (ploegh@hms.harvard.edu).

A membrane protein complex mediates retro-translocation from the ER lumen into the cytosol

Yihong Ye¹, Yoko Shibata¹, Chi Yun², David Ron² & Tom A. Rapoport¹

¹Howard Hughes Medical Institute and Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA

²Skirball Institute, New York University School of Medicine, New York, New York 10016, USA

Elimination of misfolded proteins from the endoplasmic reticulum (ER) by retro-translocation is an important physiological adaptation to ER stress. This process requires recognition of a substrate in the ER lumen and its subsequent movement through the membrane by the cytosolic p97 ATPase. Here we identify a p97-interacting membrane protein complex in the mammalian ER that links these two events. The central component of the complex, Derlin-1, is a homologue of Der1, a yeast protein whose inactivation prevents the elimination of misfolded luminal ER proteins. Derlin-1 associates with different substrates as they move through the membrane, and inactivation of Derlin-1 in *C. elegans* causes ER stress. Derlin-1 interacts with US11, a virally encoded ER protein that specifically targets MHC class I heavy chains for export from the ER, as well as with VIMP, a novel membrane protein that recruits the p97 ATPase and its cofactor.

Many proteins of eukaryotic cells undergo folding and modification in the lumen of the endoplasmic reticulum (ER). Properly folded polypeptides leave the ER along the secretory pathway, whereas misfolded proteins or unassembled protein complexes are retained¹. These proteins are eventually degraded by the proteasome and must therefore be transported back into the cytosol by a multi-step process called retro-translocation, dislocation, or ERAD (for ER-associated protein degradation) (for review, see ref. 2). Blocking retro-translocation induces the unfolded protein response (UPR), a collection of signalling pathways that adapt cells to ER stress. The retro-translocation pathway has been co-opted by certain viruses to selectively destroy cellular proteins required for the immune defence of the host. For example, the US11 protein of human cytomegalovirus (HCMV) targets newly synthesized major histocompatibility complex (MHC) class I heavy chains for retro-translocation³. These observations, and the role played by ER stress in the pathophysiology of important human diseases⁴, emphasize the need for an understanding of the mechanism of retro-translocation.

One of the best established components of the retro-translocation machinery is the cytosolic ATPase p97 (also called VCP or, in yeast, Cdc48), which interacts with a cofactor complex consisting of Ufd1 and Npl4 (ref. 5). A retro-translocation substrate emerging on the cytosolic side of the ER membrane is poly-ubiquitinated and recognized by p97 (ref. 6). The ATPase probably then 'pulls' the substrate out of the ER membrane by a mechanism similar to that proposed for AAA proteases in mitochondria and bacteria (for review, see ref. 7). In yeast, mutations in constituents of the ATPase complex block the degradation of all misfolded ER substrates tested^{8–12}. In mammals, the complex is required for the US11-mediated retro-translocation of MHC class I heavy chains^{6,9}.

Other steps in retro-translocation are less well understood. Substrate recognition must occur in the ER lumen. It appears that US11 binds specifically to MHC class I heavy chains¹³ and that certain ER chaperones recognize misfolded proteins^{12,14–18}, but how these substrates are subsequently targeted to the retro-translocation machinery and cross the ER membrane is unclear. Some evidence suggests that they may be transported through the Sec61 channel, which is responsible for transporting proteins from the cytosol into the ER, but the data are inconclusive (for discussion, see ref. 2). Importantly, there is as yet no link between substrate recognition in

the ER lumen and the function of p97 in the cytosol. Here we report the identification of a membrane protein complex that provides this link.

A membrane receptor for p97

Previous experiments had shown membrane association of p97 and its cofactor Ufd1/Npl4 and suggested the existence of a membrane receptor⁶. To identify the receptor, we first demonstrated that purified, recombinant p97 associates and co-sediments with isolated ER membranes from dog pancreas (Fig. 1a, lane 2). A p97 mutant lacking the amino-terminal domain (p97ΔN)⁶ interacted only weakly (lane 4). The membrane interaction of p97 is not mediated by its cofactor Ufd1/Npl4. After washing the membranes with high salt, Ufd1 was removed, while ~30% of the endogenous p97 remained membrane-bound (Fig. 1a, lane 10 versus 9). The cofactor-depleted membranes still bound recombinant p97 (lane 6) and again the p97ΔN had a much reduced affinity (lane 8). Together, these results indicate that p97 binds directly to ER membranes, and that high affinity binding requires its N-domain. Protease pretreatment of the salt-washed membranes abolished the interaction (data not shown), suggesting that p97 binds to an integral membrane protein receptor.

We next labelled p97 with an amino-reactive biotin derivative, containing a disulphide bridge between the reactive group and biotin. The modified p97 protein was added to salt-washed ER membranes, unbound material was removed by sedimentation, and the membranes were solubilized in digitonin. The detergent extract was incubated with streptavidin beads and bound proteins were eluted by reducing the disulphide linker. The Coomassie blue-stained gel showed two bands that were absent from the input p97 preparation (Fig. 1b, lane 2 versus 3). Because staining intensity is approximately proportional to molecular mass, it seems that the two proteins are present in near-stoichiometric amounts compared with membrane-bound p97. These proteins were not seen when p97ΔN was used, even though a small amount of p97ΔN bound to the membrane (lane 1). Sequencing by mass spectrometry gave several peptides that allowed unambiguous identification of the proteins (Supplementary Table 1).

The upper band corresponds to a homologue of *S. cerevisiae* Der1 (for 'degradation in the ER'), a protein identified in a genetic screen

for components required for the degradation of a misfolded luminal ER protein¹⁹. Essentially nothing is known about Der1's function, except that it is required for the degradation of other substrates²⁰, and that its expression is upregulated under ER stress²¹. We call the Der1 homologue Derlin-1 to indicate that it is Der1-like. Homologues of Derlin-1 are found in every eukaryotic organism (Fig. 1c; Supplementary Fig. S1). All species contain at least one other related protein, called Derlin-2, that belongs to a distinct group (Fig. S1). It is unclear which of the two classes of Der1 homologues is more closely related to yeast Der1 (E. Hartmann, personal communication). Mammalian Derlin-1 is predicted to have four transmembrane segments with both the amino and carboxy termini in the cytosol (Fig. 1c), consistent with experimental results on yeast Der1 (ref. 22).

The smaller p97-interacting protein does not have identifiable homologues outside vertebrates. It had previously turned up in a screen for selenocysteine-containing proteins²³, and is predicted to span the membrane once, with a short luminal segment and a longer cytosolic domain of ~132 amino acids (Supplementary Fig. S2). On the basis of additional data (see below), we have named this protein VIMP: for VCP (another name for p97)-interacting membrane protein (Gene bank ID: AY618665).

To confirm the interactions between Derlin-1, VIMP, and p97, we used peptide-specific antibodies to human Derlin-1 (hDerlin-1) and VIMP (see Supplementary Fig. S3). A detergent extract of HeLa cell membranes was subjected to immunoprecipitation with VIMP antibodies. Immunoblots showed that endogenous p97 and Derlin-1

were co-precipitated (Fig. 1d, lanes 2, 3). No precipitation was observed with a pre-immune serum (lane 1) or with an antiserum depleted of VIMP antibodies (lane 4). When Myc-tagged VIMP and p97 were co-expressed in HEK293 cells, the two proteins could be co-immunoprecipitated with Myc antibodies (Supplementary Fig. S4). A similar result was obtained with p97AA, a mutant defective in substrate binding⁶, indicating that VIMP does not serve as substrate for p97 (Supplementary Fig. S4). p97ΔN did not co-precipitate with Myc-VIMP. Together, these data suggest that Derlin-1 and VIMP form a membrane protein complex that serves as a receptor for p97.

VIMP links Derlin-1 with the p97 ATPase complex

Given that VIMP has a sizeable cytosolic domain, we suspected that it may be responsible for the interaction with p97. Indeed, a purified fusion protein containing glutathione S-transferase (GST) and the cytosolic domain of VIMP (GST-VIMPC) bound recombinant p97 (data not shown). Interestingly, when GST-VIMPC was added to a solution of p97, a complex containing both proteins precipitated (Fig. 2a, lanes 7–12). His-tagged VIMP also precipitated p97 (data not shown). Because both p97 and VIMP are able to form oligomers (ref. 24, and Supplementary Fig. S5), precipitation is probably caused by extensive multivalent interactions, analogous to those between a polyclonal antibody and its antigen.

To test whether VIMP can bind p97 together with its cofactor Ufd1/Npl4, we added GST-VIMPC to rat liver cytosol and analysed the resulting precipitate by immunoblotting. Indeed, the pellet

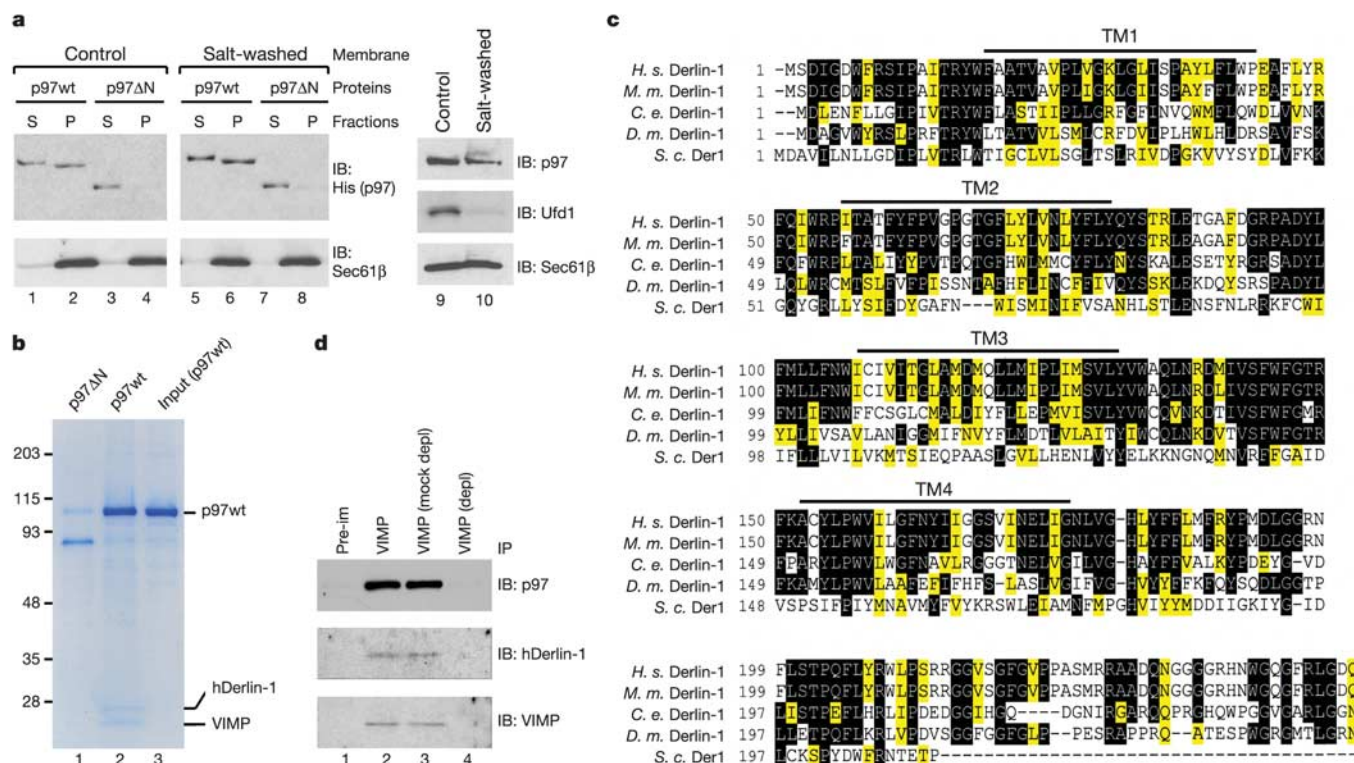


Figure 1 Identification of the Derlin-1/VIMP complex. **a**, Microsomes, washed with low (control) or high salt, were incubated with purified His-tagged wild-type p97 (p97wt) or p97 lacking the N-domain (p97ΔN). The membranes were sedimented, and the pellet (P) and supernatant (S; 20%) fractions were analysed by immunoblotting (IB) with His and Sec61β antibodies. p97 in the supernatant runs a little more slowly because of the presence of sucrose. Endogenous p97, Ufd1 and Sec61β were detected by immunoblotting (right panel). **b**, Biotinylated p97wt or p97ΔN were incubated with salt-washed microsomes. A membrane extract was incubated with streptavidin beads, and bound proteins were analysed by SDS-PAGE and Coomassie staining. An aliquot of p97wt

was loaded in parallel (Input). **c**, Sequence alignment of some Der1/Derlin homologues (*H. s.*, *Homo sapiens*; ID: NP_077271; *M. m.*, *Mus musculus*; ID: NP_077169; *C. e.*, *Caenorhabditis elegans*; ID: NP_492721; *D. m.*, *Drosophila melanogaster*; ID: NP_608632; *S. c.*, *Saccharomyces cerevisiae*). TM1 to TM4, predicted trans-membrane segments. **d**, A HeLa cell membrane extract was subjected to immunoprecipitation (IP) with preimmune (Pre-im) or VIMP antisera (depl, serum depleted of VIMP antibodies by GST-VIMPC; mock depl, mock-depleted), followed by immunoblotting with the indicated antibodies.

contained both p97 and the cofactor Ufd1 (Fig. 2b, lane 4), but not other abundant cytosolic proteins, such as proteasome subunits, Mss1 and $\alpha 7$, or Hsp70 (lanes 6–8). To test whether precipitation of the cofactor depends on p97, we depleted endogenous p97 from rat liver cytosol, using beads containing the p97-binding domain of Ufd1; this procedure removes p97, but not its cofactors, as shown by immunoblotting (Fig. 2c, lane 1). GST-VIMPC did not precipitate Ufd1 from the depleted extract (lane 4), but re-addition of p97 restored the interaction (lane 5). As expected, re-addition of the p97 Δ N mutant did not result in precipitation of the cofactor but, as before (Fig. 1b), a weak interaction with VIMP was observed (lane 6). These results show that the cytosolic domain of VIMP can bind the entire p97 ATPase complex.

Next we tested by immunofluorescence microscopy the interactions of VIMP, p97, and Derlin-1 in COS cells transiently expressing tagged versions of the proteins. When expressed individually,

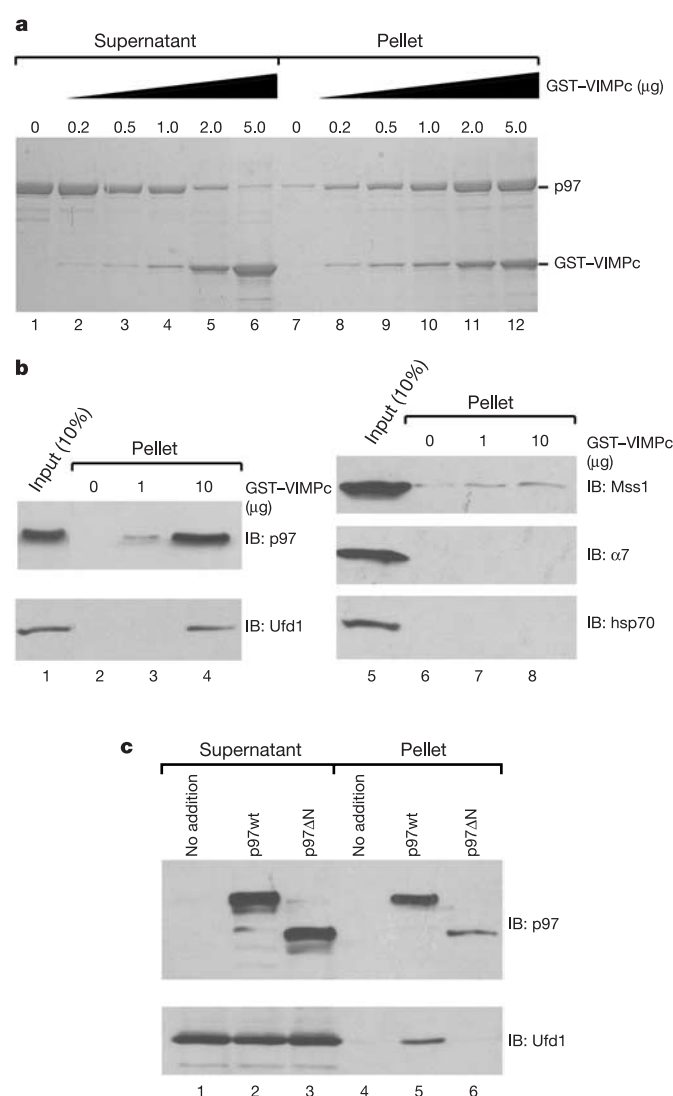


Figure 2 *In vitro* association of VIMP with p97 complexes. **a**, Purified p97 (80 nM) was incubated with the indicated amounts of a GST fusion to the cytosolic domain of VIMP (GST-VIMPC). Following centrifugation, supernatant and pellet fractions were analysed by SDS-PAGE and Coomassie staining. **b**, As in **a**, except that rat liver cytosol (250 μg) was used and the analysis was performed by immunoblotting with the indicated antibodies. **c**, As in **b**, except that cytosol depleted of p97 was used. Where indicated, p97 (p97wt) or p97 mutant (p97 Δ N) were added back to depleted cytosol (at 80 nM), before addition of GST-VIMPC.

HA-tagged hDerlin-1 (hDerlin-1-HA) showed a reticular pattern typical of ER localization, and a variable number of dots, presumably caused by its aggregation (Fig. 3a, panel 1, Supplementary Fig. S6). Myc-tagged VIMP (Myc-VIMP) was also often reticular (Fig. 3a, panel 2), but in a significant number of cells, its overexpression altered ER morphology, resulting in long filaments (Fig. 3b, panel 8, Supplementary Fig. S7). His-tagged p97 (His-p97) was localized throughout the cell (Fig. 3a, panel 3).

When Myc-VIMP and His-p97 were expressed together, the staining pattern of both proteins was changed; they co-localized in large punctae around the nucleus (Fig. 3b, panels 1–3). Although the re-localization of VIMP and p97 is probably the result of their overexpression, it demonstrates that they interact with one another. Consistent with the co-immunoprecipitation results (Supplementary Fig. S4), expression of Myc-VIMP with His-p97AA (Fig. 3b, panels 4–6), but not p97 Δ N (panels 7–9), gave rise to punctae. When hDerlin-1-HA was expressed together with Myc-VIMP and His-p97, it also relocalized to these structures (Fig. 3c, panels 1–3). In contrast, other ER membrane proteins expressed at about the same level, such as ATF6, and Sec61 β , were not recruited to them (Fig. 3c, panels 4–6, and data not shown). When Myc-VIMP was omitted, hDerlin-1 and p97 localized as when expressed alone (Fig. 3c, panels 7–9). Together, these results suggest that VIMP recruits p97 to Derlin-1, although our data do not exclude an additional interaction between p97 and Derlin-1.

Derlin-1/VIMP is involved in US11-induced retro-translocation

Next we asked whether the Derlin-1/VIMP complex is involved in retro-translocation. We first tested whether it associates with retro-translocating MHC class I heavy chains in US11-expressing astrocytoma cells. The cells were pulse-labelled with 35 S-methionine, permeabilized with digitonin, and chase-incubated in the presence of a GST-ubiquitin fusion protein (GST-Ub). The modification of retro-translocating substrates with poly(GST-Ub), instead of poly-ubiquitin, prevents their release from the ER membrane, resulting in a translocation intermediate²⁵. The membrane fraction was solubilized and subjected to sequential immunoprecipitation with Derlin-1 and heavy chain (HC) antibodies. A significant amount of heavy chains (5–10%) was found associated with Derlin-1 (Fig. 4a, lane 2), while little was precipitated with a pre-immune serum (lane 1). The immunoprecipitated material co-migrated in SDS-gels with glycosylated heavy chains (data not shown), indicating that the Derlin-1-interacting substrate had not yet reached the cytosolic N-glycanase²⁶.

Retro-translocating heavy chains could also be immunoprecipitated with VIMP antibodies. In this case, the chase incubation was carried out in the presence of a p97 mutant (p97KA) defective in ATP hydrolysis to accumulate retro-translocation intermediates on the membrane⁶. Sequential immunoprecipitation showed that increasing amounts of heavy chains became associated with VIMP during the chase period in US11-expressing cells (Fig. 4c, lanes 4–6), while little was seen in control cells (lanes 1–3). When wild-type p97 instead of p97KA was added during the chase, less substrate precipitated with VIMP antibodies (data not shown), consistent with the assumption that retro-translocating polypeptides only transiently interact with Derlin-1/VIMP.

Because the US11 protein initiates the retro-translocation of MHC class I heavy chains, we tested whether it can interact with Derlin-1. Sequential immunoprecipitation experiments with Derlin-1 and US11 antibodies did indeed show an association (Fig. 4a, lane 4 versus 3). Similar results were obtained when US11 was transiently expressed in HeLa cells; again US11 co-precipitated with Derlin-1 (Fig. 4b, lane 1). A US11 mutant (Q192L) that binds MHC class I heavy chains but cannot target them for retro-translocation²⁷ showed a reduced affinity for Derlin-1 (lane 2), indicating that the US11-Derlin-1 interaction is functionally important. An interaction between these proteins could also be seen by immunofluor-

escence microscopy in cells overexpressing Derlin-1 and US11; they co-localized in dotted structures (Supplementary Fig. S6).

US11 stably expressed in astrocytoma cells could also be immunoprecipitated with VIMP antibodies (Fig. 4d, lanes 6 and 7), whereas a pre-immune serum or an antiserum depleted of VIMP antibodies precipitated only small amounts (lanes 5 and 8). No labelled protein was precipitated from control cells that did not express US11 (lanes 1–4). The US11–VIMP interaction is probably mediated by Derlin-1, because US11 was not recruited to the punctate structures formed by co-expression of VIMP and p97 (Supplementary Fig. S6). Taken together, these results indicate that Derlin-1/VIMP associates with US11 and retro-translocating MHC class I heavy chains.

A general role for Derlin-1/VIMP in ER protein degradation

To investigate whether Derlin-1/VIMP has a general function in the retro-translocation of misfolded ER proteins, we tested its association with substrates. Pulse-labelled HeLa cells were permeabilized with digitonin, and chase-incubated in the presence of GST–Ub fusion protein. The membrane fraction was solubilized and subjected to sequential immunoprecipitation with Derlin-1 and ubiquitin antibodies. Poly(GST–Ub) modified proteins were co-precipitated with Derlin-1 under native conditions, but not after denaturation in SDS (Fig. 5a, lane 2 versus 4), although the antibodies work even better under denaturing conditions (lower panel). No precipitation was seen with the pre-immune serum (lanes 1 and 3). The association between poly(GST–Ub)-modified proteins and the Derlin-1/VIMP complex could also be seen by immunoprecipitation with VIMP antibodies under native conditions (Fig. 5b, lanes 2 and 3). Much less material was precipitated with pre-immune serum (lane 1) or an antiserum depleted of VIMP antibodies (lane 4). When HeLa cells were treated during the pulse-labelling period with dithiothreitol (DTT) to increase the amount of misfolded ER proteins²⁸, more poly(GST–Ub) modified material was found associated with VIMP (Fig. 5b, lanes 6, 7 versus lanes 2, 3). Together, these results suggest that Derlin-1/VIMP associates

with a large set of substrates undergoing retro-translocation.

If Derlin-1 plays a general role in retro-translocation, one would expect that its depletion by RNA interference (RNAi) would lead to the accumulation of misfolded proteins in the ER and elicit the UPR. To test this assumption, we used a strain of *C. elegans*, which expresses the green fluorescent protein (GFP) under the control of the UPR-inducible *hsp4* promoter (*hsp4::gfp*)²⁹. Inactivation of *C. elegans* Derlin-1 (cDerlin-1) by RNAi activated the UPR reporter in many different cells of the animal, as demonstrated by fluorescence microscopy (Fig. 5c, panel 2 versus 1) and immunoblotting for GFP (Fig. 5d, lane 2 versus 1). The expression of *hsp4::gfp* was blocked in mutant *ire1* animals (Fig. 5c, panel 3), as predicted from the essential role of the Ire1 kinase in UPR²⁹. The level of *hsp4::gfp* expression by cDerlin-1-depletion was comparable to that observed after inactivation of Sel1 (Fig. 5c, panel 2 versus 4; Fig. 5d, lane 2 versus 4), which encodes a homologue of yeast Hrd3 involved in the poly-ubiquitination of a subset of retro-translocation substrates^{30,31}. RNAi-mediated depletion of Ero1, a protein required for disulphide bond formation in the ER^{32,33}, led to an even stronger activation of UPR (Fig. 5c, panel 5; Fig. 5d, lane 5). The lower level of *hsp4::gfp* activation observed following cDerlin-1 or Sel1 depletion is consistent with their function in removing misfolded ER proteins under physiological conditions, rather than in serving as general folding catalysts. As in yeast^{19,30,31}, neither cDerlin-1 nor Sel1 is essential for viability. Depletion of Derlin-2 (cDerlin-2; Supplementary Fig. S1) in wild-type animals did not elicit the UPR (Fig. 5c, panel 6; Fig. 5d, lane 3), suggesting that it may not be involved in retro-translocation or may be required for the degradation of a restricted subset of substrates.

Discussion

The discovery of the Derlin-1/VIMP membrane protein complex in mammalian cells provides a link between substrate recognition in the ER and the function of the p97 ATPase in the cytosol. It explains how the p97 complex is recruited to the ER membrane for retro-translocation, and suggests a mechanism by which the viral US11

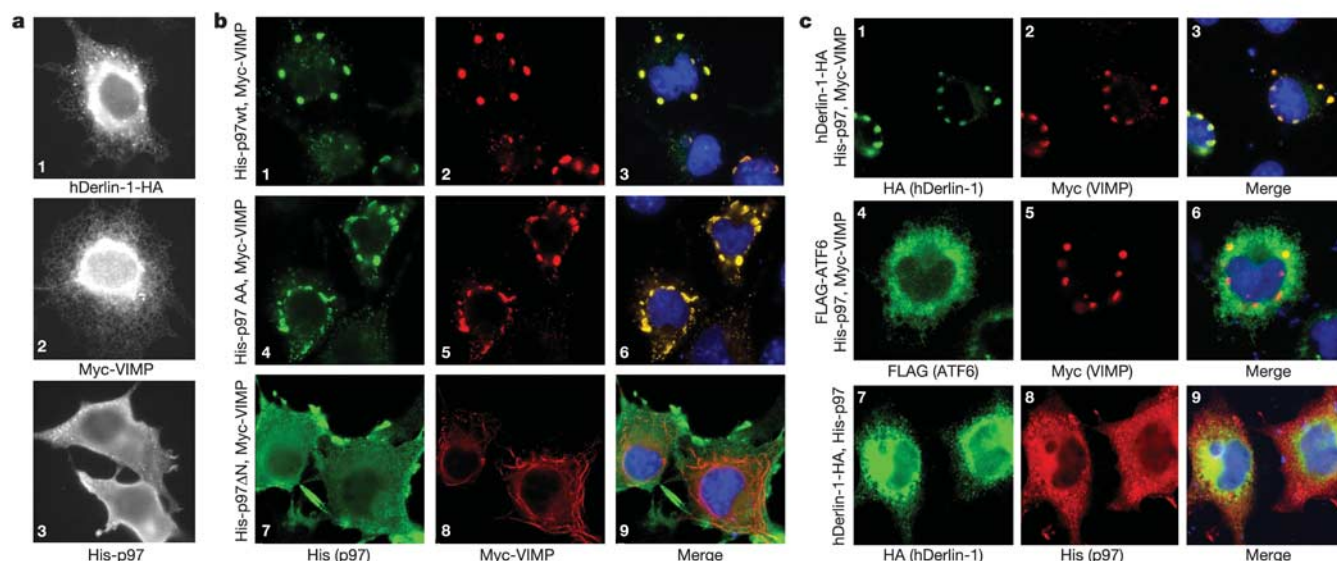


Figure 3 VIMP mediates p97 binding to hDerlin-1. **a**, HA-tagged Derlin-1 (hDerlin-1-HA, panel 1), Myc-tagged VIMP (Myc-VIMP, panel 2), or His-tagged p97 (His-p97, panel 3) were expressed individually in COS cells, stained with the corresponding antibodies, and visualized by fluorescence microscopy. **b**, Myc-VIMP was expressed together with His-tagged p97 (His-p97wt), p97 lacking the N-domain (His-p97ΔN), or p97 defective in ATP and substrate binding (His-p97AA). The cells were stained with His or Myc antibodies, as

indicated. Panels 3, 6 and 9 were also stained with 4,6-diamidino-2-phenylindole (DAPI) and show merged images. **c**, His-p97wt, Myc-VIMP, and hDerlin-1-HA were coexpressed (panels 1–3). Controls were performed by replacing hDerlin-1-HA with FLAG-tagged ATF6 (FLAG-ATF6, panels 4–6), or by omitting Myc-VIMP (panels 7–9). The cells were stained with HA, FLAG or Myc antibodies, as indicated. Panels 3, 6 and 9 show merged images.

protein co-opts the cellular pathway for destruction of MHC class I heavy chains.

Derlin-1/Der1 appears to be a central, evolutionarily conserved membrane component of the retro-translocation machinery. It is required for the degradation of different misfolded substrates in yeast^{19,20}, it is associated with substrates that are retro-translocating through the mammalian ER membrane, its depletion in *C. elegans* results in ER stress, and its expression is upregulated during ER stress in yeast²¹ and *C. elegans* (D.R., unpublished results). A functional role for VIMP in retro-translocation is suggested by its association with Derlin-1 and retro-translocating substrates, by its ability to bind p97 and its cofactor, and by its upregulation under ER stress conditions³⁴, but direct evidence remains to be provided. Because Derlin-1 can interact with US11 and VIMP, factors that perform luminal and cytosolic functions, respectively, we propose that it is a central component of a retro-translocation channel.

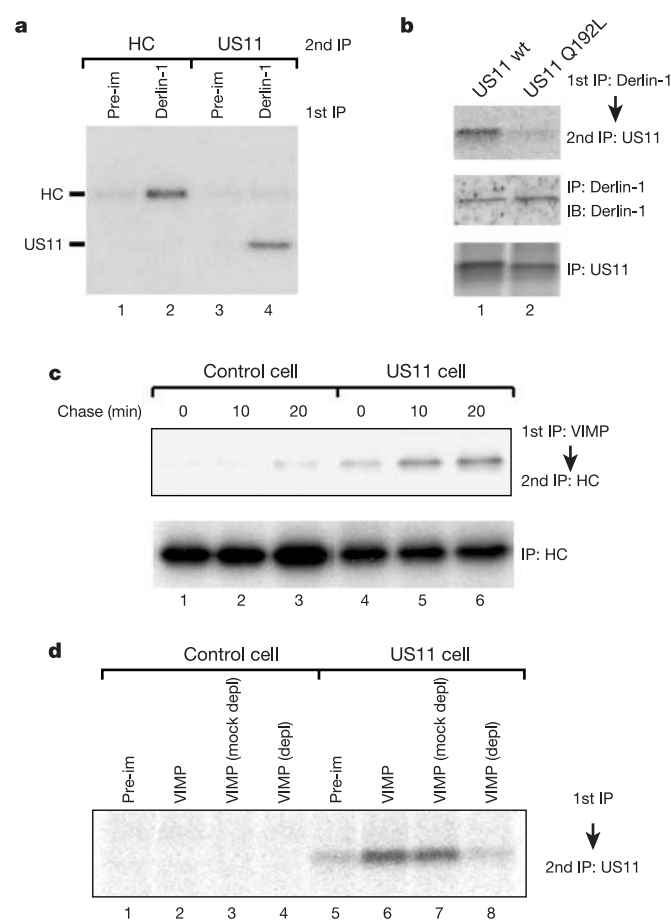


Figure 4 Derlin-1/VIMP associates with MHC class I heavy chains and US11. **a**, US11-expressing astrocytoma cells (US11) were labelled with ³⁵S-methionine, permeabilized, and chase-incubated in the presence of GST-ubiquitin. A membrane extract was subjected to immunoprecipitation with pre-immune or Derlin-1 antiserum (1st IP), followed by immunoprecipitation with HC or US11 antibodies (2nd IP). The samples were analysed by SDS-PAGE and autoradiography. **b**, A cell extract of radiolabelled HeLa cells expressing wild-type or mutant US11 (US11 wt or Q192L) was subjected to sequential immunoprecipitation with Derlin-1 and US11 antibodies (top panel). A portion of the Derlin-1 immunoprecipitate was analysed by immunoblotting with Derlin-1 antibodies (middle panel), and a direct immunoprecipitation with US11 antibodies was performed in parallel (lower panel). **c**, Radiolabelled, permeabilized US11 cells or control cells were chase-incubated in the presence of mutant p97 (p97KA) for different time periods. A membrane extract was subjected to sequential immunoprecipitation with VIMP and HC antibodies (upper panel) or to direct immunoprecipitation with HC antibodies (lower panel). **d**, As in **c**, with the sequential immunoprecipitation performed with the indicated antibodies.

A complete picture for the US11-induced retro-translocation of MHC class I heavy chains, from recognition of the substrate in the ER to its degradation by the proteasome in the cytosol, is now emerging (Fig. 6). First, US11 interacts with heavy chains in the ER lumen and targets them to Derlin-1, as is also reported in a parallel paper⁴⁰. US11 might cycle on and off Derlin-1 and deliver the substrate in a catalytic manner, perhaps with the help of additional components. Next, the substrate is translocated across the membrane through a protein-conducting channel that is postulated to contain Derlin-1 (Fig. 6). Once a segment of the heavy chain has emerged into the cytosol, it is captured by the p97 ATPase complex, and at the same time, poly-ubiquitin chains are attached. Finally, p97 and its cofactor Ufd1/Npl4 pull the substrate out of the ER for proteasomal degradation.

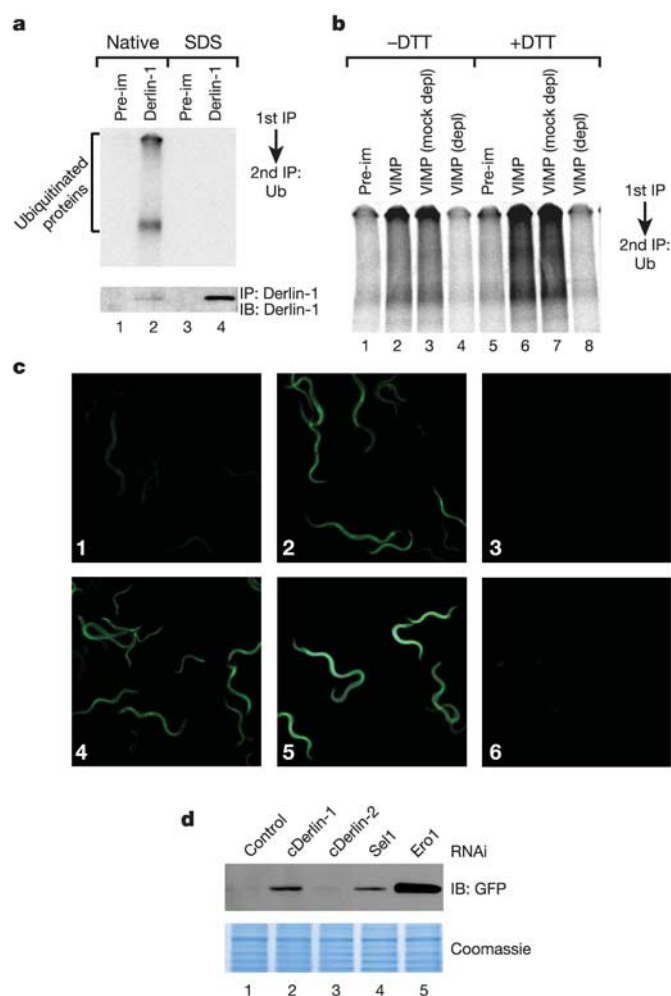


Figure 5 Derlin-1/VIMP associates with misfolded proteins and Derlin-1 depletion causes ER stress. **a**, Radiolabelled, permeabilized HeLa cells were chase-incubated in the presence of GST-ubiquitin. Membrane extracts made in a mild detergent (native) or SDS (denaturing) were subjected to immunoprecipitation with preimmune or Derlin-1 antiserum, followed by immunoprecipitation with ubiquitin (Ub) antibodies (upper panel). A portion of the first immunoprecipitate was analysed by immunoblotting with Derlin-1 antibodies (lower panel). **b**, As in **a**, except that the first immunoprecipitation was performed with the indicated VIMP antisera. Where indicated, DTT (5 mM) was included during pulse-labelling (5 min). **c**, Wild-type *C. elegans* animals (panels 1, 2 and 4–6) or mutant animals lacking Ire1 (*ire1(zc14)II*) (panel 3), all expressing GFP under an UPR-inducible promoter (*hsp4::gfp(zcls4)V*), were fed with different RNAi constructs to deplete the following proteins: panel 1, no depletion (control plasmid); panels 2 and 3, cDerlin-1; panel 4, Sel1; panel 5, Ero1; panel 6, cDerlin-2. **d**, Extracts of different animals were analysed by immunoblotting with GFP antibodies (top panel) or Coomassie staining (bottom panel).

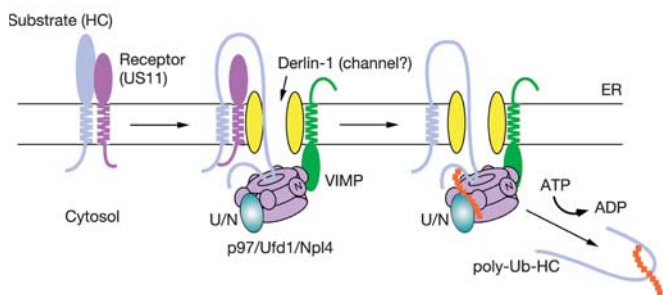


Figure 6 Model for US11-mediated retro-translocation of MHC class I heavy chains. US11 recognizes HC in the ER lumen and targets it to Derlin-1, a proposed component of the retro-translocation channel. The p97 ATPase complex is recruited to Derlin-1 by VIMP. HC emerging into the cytosol is bound by p97. Poly-ubiquitin chains (Poly-Ub, red) are attached and recognized by both the N-domain (N) of p97 and the cofactor Ufd1/Npl4 (U/N). ATP hydrolysis by p97 moves HC into the cytosol. The retro-translocation of misfolded ER proteins may occur similarly, with US11 being replaced by other targeting components.

The retro-translocation of misfolded ER proteins probably occurs by the same pathway, the major difference being in substrate recognition. Initial recognition may be mediated by certain ER chaperones, such as protein disulphide isomerase or calnexin/calreticulin^{12,14,16–18}, but targeting to Derlin-1 may require additional shuttle proteins that can interact with both the chaperone-associated substrate and Derlin-1/VIMP. The viral US11 protein appears to combine both functions. The retro-translocation of misfolded proteins in lower organisms also requires a different ‘cytosolic adaptor’ to recruit the p97 ATPase, because homologues of VIMP appear to exist only in vertebrates.

Experiments in yeast show that Der1 is only required for the degradation of ER proteins with misfolded luminal domains²⁰. Membrane proteins with misfolded cytosolic or intra-membrane domains do not require Der1 and might move through a different translocation channel, perhaps the Sec61 channel. Alternatively, they may not need a channel at all, as suggested by the bacterial AAA protease FtsH, which can degrade a misfolded membrane protein in a purified reconstituted system lacking any obvious channel³⁵. All pathways of retro-translocation, whether Der1/Derlin-1-dependent or not, appear to require the function of the p97 ATPase complex, which may provide the general driving force for the movement of proteins into the cytosol. □

Methods

Constructs

Constructs that express His-tagged versions of wild type p97 (His-p97wt), a mutant lacking the N domain (His-p97ΔN), and a mutant defective in ATP and substrate binding (His-p97AA) in *E. coli* were described previously⁶. To generate mammalian expression constructs containing these p97 variants, the coding regions in the bacterial expression plasmids were amplified by polymerase chain reaction (PCR) and inserted into the pcDNA3.1 vector using the TOPO cloning technology (Invitrogen). Human Derlin-1 (hDerlin-1) (Gene bank ID: NP_077271) and VIMP (Gene bank ID: AY618665) complementary DNAs were isolated from a cDNA library (Panomics) by PCR and cloned into the pcDNA3.1 vector. Sequencing of several independent PCR clones of hDerlin-1 revealed a single base variation from the reported sequence, resulting in a G to R substitution at residue 245. To generate a construct expressing a GST fusion to the cytosolic domain of VIMP (GST-VIMPC), a DNA segment encoding amino acids 49 to 187 of VIMP was inserted into the pET42 vector (Novagen). The RNAi constructs for *C. elegans* Derlin-1 homologues were generated by cloning a cDNA segment encoding residues 68 to 242 of cDerlin-1 (F25D7.1; ID: NP_492721) and a segment encoding residue 24 to 221 of cDerlin-2 (R151.6; ID: T16766) into the pPD129.36 vector³⁶. The Sel1 and Ero1 RNAi constructs were described^{37,38}. The plasmid pCMV-Flag-ATF6 was provided by R. Prywes. The plasmid pCMV-US11 was described previously³. The US11 Q192L mutant was made by site-directed mutagenesis.

Antibodies and proteins

Antibodies to MHC class I heavy chain, Sec61β, p97, His-tag, and ubiquitin were described previously⁶. Antibodies to mammalian Ufd1 were described⁵. Myc antibodies

(9E10), HA antibodies (12CA5), and FLAG M2 antibodies were purchased from Santa Cruz, Roche and Sigma, respectively. Antibodies to proteasome subunits (α7 and Mms1), to the cytosolic chaperone Hsp70, and to GFP were from Affiniti Bioreagents, StressGen Biotechnology and Molecular Probes, respectively. Peptide-specific antibodies against hDerlin-1 and VIMP were raised in rabbits. Two peptides corresponding to residues 204 to 216 and 238 to 251 of hDerlin-1, and a peptide corresponding to residues 174 to 187 of human VIMP were chosen by Zymed for immunization.

The purification of His-tagged p97 variants was described in ref. 6. GST-VIMPC was expressed in *E. coli* and purified with glutathione beads. The eluted protein was further purified on a Superdex 200 HR (10/30) column in 50 mM Tris/HCl pH 8.0, 150 mM potassium chloride, 5% glycerol and 2 mM magnesium chloride. His-VIMPC was expressed in *E. coli* and purified by Ni-NTA chromatography. GST-ubiquitin was purchased from BostonBiochem.

Membrane binding assay

Purified p97 variants (250 nM) were incubated with 60 equivalents of dog pancreatic microsomes for 20 min at 4°C. The membranes were layered on top of a sucrose cushion (50 mM HEPES pH 7.4, 150 mM potassium acetate, 10 mM magnesium acetate, 1.2 M sucrose) and sedimented by centrifugation in a TLA100 rotor for 15 min at 75,000 r.p.m. Supernatant fractions were analysed directly, while the pellets were washed with membrane buffer (50 mM HEPES pH 7.4, 150 mM potassium acetate, 10 mM magnesium acetate and 250 mM sucrose) before analysis by SDS-polyacrylamide gel electrophoresis (PAGE) and immunoblotting. Where indicated, peripheral membrane proteins were removed from microsomes by washing the membranes twice with membrane buffer containing 800 mM sodium chloride. Salt-washed membranes were resuspended in membrane buffer and stored at –80°C.

Purification of Derlin-1/VIMP complex

Biotinylation of purified p97 variants was performed using EZ-link sulfo-NHS-SS-Biotin (Pierce) according to the instructions of the manufacturer. Modified p97 or p97ΔN (100 μg) was incubated with 6,000 equivalents salt-washed dog pancreatic microsomes for 30 min at 4°C. Unbound proteins were removed by sedimentation of the membranes through a sucrose cushion. Membranes were solubilized in 50 mM HEPES pH 7.4, 200 mM potassium acetate and 10 mM magnesium, 2.5% digitonin plus protease inhibitor cocktail. The extract was incubated with streptavidin beads and bound p97 was eluted in solubilization buffer containing 50 mM DTT. The proteins were subjected to SDS-PAGE and Coomassie staining. Two visible bands were cut out and analysed by mass spectrometry. Sequences of five peptides from the 28-kD bands and three peptides from the 25-kD bands were obtained (Supplementary Table 1). These peptides match the sequences of two proteins in the human genome sequence database (ID: NP_077271 and NP_060915, respectively).

In vitro co-precipitation experiments

To determine the association of VIMP with p97 and its cofactor, the indicated amount of GST-VIMPC was incubated with either 80 nM purified p97 or with 100 μl rat liver cytosol. The samples were centrifuged for 20 min at 20,000 g, and the resulting supernatant and pellet fractions were analysed by SDS-PAGE, followed by Coomassie staining or immunoblotting.

Mammalian cell culture and transient gene expression

HeLa cells, HEK293 cells, and COS cells were maintained according to standard procedures. Transient transfections were performed either following the standard calcium phosphate precipitation protocol for HEK293 cells, or using the FuGENE 6 reagent (Roche) for HeLa and COS cells. For immunofluorescence microscopy, cells were fixed with 3% paraformaldehyde and stained with antibodies to Myc (1:1000), His (1:100), HA (1:1000), FLAG (1:1000) or US11 (1:2000).

Association of Derlin-1/VIMP with retro-translocating substrates and US11

The use of permeabilized US11-expressing astrocytoma cells to study the retro-translocation of MHC class I heavy chains was described³⁹. To analyse the association of Derlin-1/VIMP with retro-translocating HC, ³⁵S-methionine labelled cells were permeabilized and chase-incubated for 40 min in the presence of either 20 μM GST-ubiquitin or 300 nM mutant p97 protein (p97KA)^{6,25}. The membranes were solubilized in buffer N (1% DeoxyBig CHAP, 30 mM Tris/HCl pH 7.4, 150 mM potassium acetate, 4 mM magnesium acetate, 1 mM ATP and protease inhibitor cocktail), and extracts were subjected to immunoprecipitation with antibodies to Derlin-1 or VIMP, followed by immunoprecipitation with HC antibodies. To determine the interaction of Derlin-1/VIMP with poly-ubiquitinated substrates, HeLa cells were labelled with ³⁵S-methionine and permeabilized as was done with US11 cells. GST-Ub (20 μM) was included during the chase incubation. Membranes were solubilized in buffer N and subjected to immunoprecipitation with antibodies to Derlin-1 or VIMP, followed by immunoprecipitation with ubiquitin antibodies. The association of Derlin-1/VIMP with US11 was determined in a similar way, except that the second immunoprecipitation was performed with US11 antibodies.

RNAi experiments in *C. elegans*

C. elegans were maintained on nematode growing media (NGM) agar plates according to standard protocols. For RNAi experiments, 3–6 animals at L4 stage were placed on an isopropyl-β-D-thiogalactopyranoside-containing NGM plate pre-seeded with bacteria carrying an RNAi feeding construct. Phenotypes were examined three days later with a fluorescent microscope. To verify the UPR induction by immunoblotting, animals from three plates were washed off with M3 medium and homogenized in 20 mM HEPES pH 7.4,

50 mM sodium chloride, 10% glycerol, 1% Triton X-100 and 1 mM EDTA plus protease inhibitors. Extracts corresponding to 20 µg proteins were fractionated on a SDS–PAGE gel and analysed by either immunoblotting with GFP antibodies or Coomassie staining.

Received 5 March; accepted 17 May 2004; doi:10.1038/nature02656.

1. Ellgaard, L. & Helenius, A. ER quality control: towards an understanding at the molecular level. *Curr. Opin. Cell Biol.* **13**, 431–437 (2001).
2. Tsai, B., Ye, Y. & Rapoport, T. A. Retro-translocation of proteins from the endoplasmic reticulum into the cytosol. *Nature Rev. Mol. Cell Biol.* **3**, 246–255 (2002).
3. Wiertz, E. J. H. J. *et al.* The human cytomegalovirus US11 gene product dislocates MHC class I heavy chains from the endoplasmic reticulum to the cytosol. *Cell* **84**, 769–779 (1996).
4. Aridor, M. & Balch, W. E. Integration of endoplasmic reticulum signaling in health and disease. *Nature Med.* **5**, 745–751 (1999).
5. Meyer, H. H., Shorter, J. G., Seemann, J., Pappin, D. & Warren, G. A complex of mammalian ufd1 and npl4 links the AAA-ATPase, p97, to ubiquitin and nuclear transport pathways. *EMBO J.* **19**, 2181–2192 (2000).
6. Ye, Y., Meyer, H. H. & Rapoport, T. A. Function of the p97-Ufd1-Npl4 complex in retrotranslocation from the ER to the cytosol: dual recognition of nonubiquitinated polypeptide segments and polyubiquitin chains. *J. Cell Biol.* **162**, 71–84 (2003).
7. Langer, T. AAA proteases: cellular machines for degrading membrane proteins. *Trends Biochem. Sci.* **25**, 247–251 (2000).
8. Bays, N. W., Wilhovsky, S. K., Goradia, A., Hodgkiss-Harlow, K. & Hampton, R. Y. HRD4/NPL4 is required for the proteasomal processing of ubiquitinated ER proteins. *Mol. Biol. Cell* **12**, 4114–4128 (2001).
9. Ye, Y., Meyer, H. H. & Rapoport, T. A. The AAA ATPase Cdc48/p97 and its partners transport proteins from the ER into the cytosol. *Nature* **414**, 652–656 (2001).
10. Jarosch, E. *et al.* Protein dislocation from the ER requires polyubiquitination and the AAA-ATPase Cdc48. *Nature Cell Biol.* **4**, 134–139 (2002).
11. Rabinovich, E., Kerem, A., Frohlich, K. U., Diamant, N. & Bar-Nun, S. AAA-ATPase p97/Cdc48p, a cytosolic chaperone required for endoplasmic reticulum-associated protein degradation. *Mol. Cell Biol.* **22**, 626–634 (2002).
12. Wang, Q. & Chang, A. Substrate recognition in ER-associated degradation mediated by Eps1, a member of the protein disulfide isomerase family. *EMBO J.* **22**, 3792–3802 (2003).
13. Story, C. M., Furman, M. H. & Ploegh, H. L. The cytosolic tail of class I MHC heavy chain is required for its dislocation by the human cytomegalovirus US2 and US11 gene products. *Proc. Natl Acad. Sci. USA* **96**, 8516–8521 (1999).
14. Gillette, P., Luz, J. M., Lennarz, W. J., de La Cruz, F. J. & Romisch, K. Export of a cysteine-free misfolded secretory protein from the endoplasmic reticulum for degradation requires interaction with protein disulfide isomerase. *J. Cell Biol.* **147**, 1443–1456 (1999).
15. Nishikawa, S. I., Fewell, S. W., Kato, Y., Brodsky, J. L. & Endo, T. Molecular chaperones in the yeast endoplasmic reticulum maintain the solubility of proteins for retrotranslocation and degradation. *J. Cell Biol.* **153**, 1061–1070 (2001).
16. Tsai, B., Rodighiero, C., Lencer, W. I. & Rapoport, T. A. Protein disulfide isomerase acts as a redox-dependent chaperone to unfold cholera toxin. *Cell* **104**, 937–948 (2001).
17. Molinari, M., Calanca, V., Galli, C., Lucca, P. & Paganetti, P. Role of EDEM in the release of misfolded glycoproteins from the calnexin cycle. *Science* **299**, 1397–1400 (2003).
18. Oda, Y., Hosokawa, N., Wada, I. & Nagata, K. EDEM as an acceptor of terminally misfolded glycoproteins released from calnexin. *Science* **299**, 1394–1397 (2003).
19. Knop, M., Finger, A., Braun, T., Hellmuth, K. & Wolf, D. H. Der1, a novel protein specifically required for endoplasmic reticulum degradation in yeast. *EMBO J.* **15**, 753–763 (1996).
20. Vashist, S. & Ng, D. T. Misfolded proteins are sorted by a sequential checkpoint mechanism of ER quality control. *J. Cell Biol.* **165**, 41–52 (2004).
21. Travers, K. J. *et al.* Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation. *Cell* **101**, 249–258 (2000).
22. Hitt, R. & Der Wolf, D. H. Ip, a protein required for degradation of malformed soluble proteins of the endoplasmic reticulum: topology and Der1-like proteins. *FEMS Yeast Res.* **4**, 721–729 (2004).
23. Kryukov, G. V. *et al.* Characterization of mammalian selenoproteomes. *Science* **300**, 1439–1443 (2003).
24. DeLaBarre, B. & Brunger, A. T. Complete structure of p97/valosin-containing protein reveals communication between nucleotide domains. *Nature Struct. Biol.* **10**, 856–863 (2003).
25. Flierman, D., Ye, Y., Dai, M., Chau, V. & Rapoport, T. A. Polyubiquitin serves as a recognition signal, rather than a ratcheting molecule, during retrotranslocation of proteins across the endoplasmic reticulum membrane. *J. Biol. Chem.* **278**, 34774–34782 (2003).
26. Wiertz, E. J. H. J. *et al.* Sec61-mediated transfer of a membrane protein from the endoplasmic reticulum to the proteasome for destruction. *Nature* **384**, 432–438 (1996).
27. Lilley, B. N., Tortorella, D. & Ploegh, H. L. Dislocation of a type I membrane protein requires interactions between membrane-spanning segments within the lipid bilayer. *Mol. Biol. Cell* **14**, 3690–3698 (2003).
28. Braakman, L., Helenius, J. & Helenius, A. Manipulating disulfide bond formation and protein folding in the endoplasmic reticulum. *EMBO J.* **11**, 1717–1722 (1992).
29. Calfon, M. *et al.* IRE1 couples endoplasmic reticulum load to secretory capacity by processing the XBP-1 mRNA. *Nature* **415**, 92–96 (2002).
30. Plemper, R. K. *et al.* Genetic interactions of Hrd3p and Der3p/Hrd1p with Sec61p suggest a retrotranslocation complex mediating protein transport for ER degradation. *J. Cell Sci.* **112**, 4123–4134 (1999).
31. Gardner, R. G. *et al.* Endoplasmic reticulum degradation requires lumen to cytosol signaling. Transmembrane control of Hrd1p by Hrd3p. *J. Cell Biol.* **151**, 69–82 (2000).
32. Frand, A. R. & Kaiser, C. A. The ERO1 gene of yeast is required for oxidation of protein dithiols in the endoplasmic reticulum. *Mol. Cell* **1**, 161–170 (1998).
33. Pollard, M. G., Travers, K. J. & Weissman, J. S. Ero1p: a novel and ubiquitous protein with an essential role in oxidative protein folding in the endoplasmic reticulum. *Mol. Cell* **1**, 171–182 (1998).
34. Gao, Y. *et al.* Regulation of the selenoprotein SelS by glucose deprivation and endoplasmic reticulum stress—SelS is a novel glucose-regulated protein. *FEBS Lett.* **563**, 185–190 (2004).
35. Akiyama, Y. & Ito, K. Reconstitution of membrane proteolysis by FtsH. *J. Biol. Chem.* **278**, 18146–18153 (2003).
36. Timmons, L., Court, D. L. & Fire, A. Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in *Caenorhabditis elegans*. *Gene* **263**, 103–112 (2001).
37. Urano, F. *et al.* A survival pathway for *Caenorhabditis elegans* with a blocked unfolded protein response. *J. Cell Biol.* **158**, 639–646 (2002).
38. Harding, H. P. *et al.* An integrated stress response regulates amino acid metabolism and resistance to oxidative stress. *Mol. Cell* **11**, 619–633 (2003).
39. Shamu, C. E., Story, C. M., Rapoport, T. A. & Ploegh, H. L. The pathway of US11-dependent degradation of MHC class I heavy chains involves a ubiquitin-conjugated intermediate. *J. Cell Biol.* **147**, 45–58 (1999).
40. Lilley, B. N. & Ploegh, H. L. A membrane protein required for dislocation of misfolded proteins from the ER. *Nature* (this issue).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. Prywes for ATF6 plasmid, the Taplin Mass Spectrometry facility (Harvard Medical School) for protein identification, and the Nikon imaging facility (Harvard Medical School) for assistance in microscopy. We thank E. Hartmann for help with sequence analysis, D. Flierman for help with the Derlin-1 immunoprecipitation, and C. Shamu, D. Finley and K. Cannon for critical reading of the manuscript. Y.Y. is supported by a Helen Hay Whitney fellowship and T.A.R. by an NIH grant. C.Y. is supported by an NIH fellowship and D.R. by NIH grants and the Ellison Medical Foundation. T.A.R. is a Howard Hughes Medical Institute investigator.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.A.R. (tom_rapoport@mms.harvard.edu).

Mercury's capture into the 3/2 spin-orbit resonance as a result of its chaotic dynamics

Alexandre C. M. Correia^{1,2} & Jacques Laskar²

¹Departamento de Física da Universidade de Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal

²Astronomie et Systèmes Dynamiques, IMCCE-CNRS UMR8028, Observatoire de Paris, 77 Avenue Denfert-Rochereau, 75014 Paris, France

Mercury is locked into a 3/2 spin-orbit resonance where it rotates three times on its axis for every two orbits around the sun^{1–3}. The stability of this equilibrium state is well established^{4–6}, but our understanding of how this state initially arose remains unsatisfactory. Unless one uses an unrealistic tidal model with constant torques (which cannot account for the observed damping of the libration of the planet) the computed probability of capture into 3/2 resonance is very low (about 7 per cent)⁵. This led to the proposal that core–mantle friction may have increased the capture probability, but such a process requires very specific values of the core viscosity^{7,8}. Here we show that the chaotic evolution of Mercury's orbit can drive its eccentricity beyond 0.325 during the planet's history, which very efficiently leads to its capture into the 3/2 resonance. In our numerical integrations of 1,000 orbits of Mercury over 4 Gyr, capture into the 3/2 spin-orbit resonant state was the most probable final outcome of the planet's evolution, occurring 55.4 per cent of the time.

Tidal dissipation will drive the rotation rate of the planet towards a limit equilibrium value $x_1(e)n$ depending on the eccentricity e and on the mean motion n (see Methods). In a circular orbit ($e = 0$) this equilibrium coincides with synchronization ($x_1(0) = 1$), but $x_1(e_0) = 1.25685$ for the present value of Mercury's eccentricity ($e_0 = 0.206$), while the equilibrium rotation rate $3n/2$ is achieved for $e_{3/2} = 0.284927$. In their seminal work⁵, Goldreich and Peale assumed that Mercury passed through the 3/2 resonance during its initial spin-down. They derived an analytical estimate of the capture probability into the 3/2 resonance and found $P_{3/2} = 6.7\%$ for the eccentricity e_0 . With the updated value of the momentum of inertia⁹ $(B - A)/C \approx 1.2 \times 10^{-4}$, this probability increases to 7.73%, and our numerical simulations with the same setting give $P_{3/2} = 7.10\%$ with satisfactory agreement.

In fact, using the present value of the eccentricity of Mercury is questionable, as the eccentricity undergoes strong variations in time, owing to planetary secular perturbations. Assuming a random date for the crossing of the 3/2 resonance for 2,000 orbits, we found numerically $P_{3/2}^{\text{BVW50}} = 3.92\%$ and $P_{3/2}^{\text{BRE74}} = 5.48\%$ for these secular (averaged) solutions of Brouwer and Van Woerkom¹⁰ and Bretagnon¹¹. It should be stressed that with the regular quasiperiodic solutions BVW50 or BRE74, as for the fixed value of the eccentricity e_0 , the 3/2 resonance can be crossed only once, because $e < e_{3/2}$. This will no longer be the case with a complete solution for Mercury's orbit that takes into account its chaotic evolution^{12,13}. In this case, Mercury's eccentricity can exceed the characteristic value $e_{3/2}$ (Fig. 1), and additional capture into resonance can occur.

To check this new scenario, it is not possible to use a single orbital solution because, owing to its chaotic behaviour, the motion cannot be predicted precisely beyond a few tens of millions of years. We have thus performed a statistical study of the past evolutions of Mercury's orbit, with the integration of 1,000 orbits over 4 Gyr in the past, starting with very close initial conditions. This statistical study was made possible by the use of the averaged equations for the motion of the Solar System^{12,13} that have recently been readjusted and compared to recent numerical integrations¹⁴, with very good

agreement over nearly 35 Myr.

Owing to the chaotic evolution, the density function of the 1,000 solutions over 4 Gyr is a smooth function (Fig. 1), similar, but not equal, to a gaussian curve¹⁴. The mean value of the eccentricity $\bar{e}_{\text{LA04}}$ is slightly higher than $\bar{e}_{\text{BVW50}}$ and $\bar{e}_{\text{BRE74}}$, but the main difference is a much wider range for the eccentricity variations, from nearly zero to more than 0.45. The planet eccentricity can now increase beyond $e_{3/2}$ during its history. Even if these episodes do not last for a long time, they will allow additional capture into the 3/2 spin-orbit resonance.

For each of these 1,000 orbital motions of Mercury, we have numerically integrated the rotational motion of the planet, taking into account the resonant terms of equation (2), for $p = k/2$ with $k = 1, \dots, 10$, the tidal dissipation, and the planetary perturbations, starting at $t_0 = -4$ Gyr, with a rotation period of 20 days. Because e is not constant, the ratio $x(t)$ of the rotation rate of the planet to its mean motion n will tend towards a limit value $\bar{x}_1(t)$ (see Methods) that is similar to an averaged value of $x_1(e(t))$, and capture into resonance can now occur in various ways.

Type I is the classical case, where $e < e_p$ (Fig. 2a). It is only in this case that the probability formula of Goldreich and Peale⁵ will apply. In type II, the eccentricity oscillates around e_p at the time when the spin rate $x(t)$ decreases towards p . The tidal dissipation thus drives $x(t)$ several times across p , greatly increasing the probability of capture (Fig. 2b). Types I and II can only occur in the first few Myr, as the spin rate decreases from faster rotations. We distinguish these cases from type III, where the planet is not initially captured into

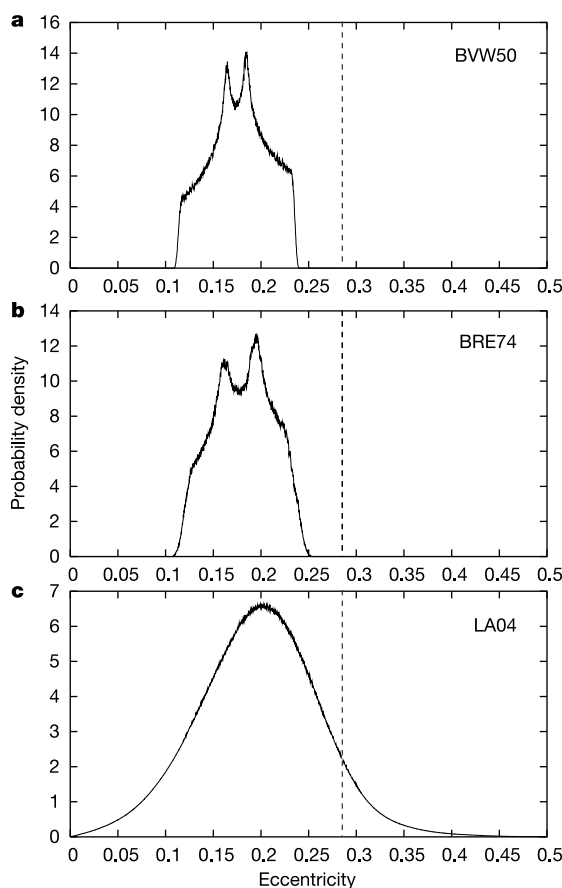


Figure 1 Probability density function of Mercury's eccentricity. Values are computed over 4 Gyr for the two quasiperiodic solutions BVW50 (ref. 10) (a) and BRE74 (ref. 11) (b) and for the numerical integration of the secular equations of refs 12 and 14 for 1,000 close initial conditions (LA04, c). The mean values of the eccentricity in these solutions are respectively $\bar{e}_{\text{BVW50}} = 0.177$, $\bar{e}_{\text{BRE74}} = 0.181$, and $\bar{e}_{\text{LA04}} = 0.198$. The vertical dotted line is the characteristic value $e_{3/2}$.

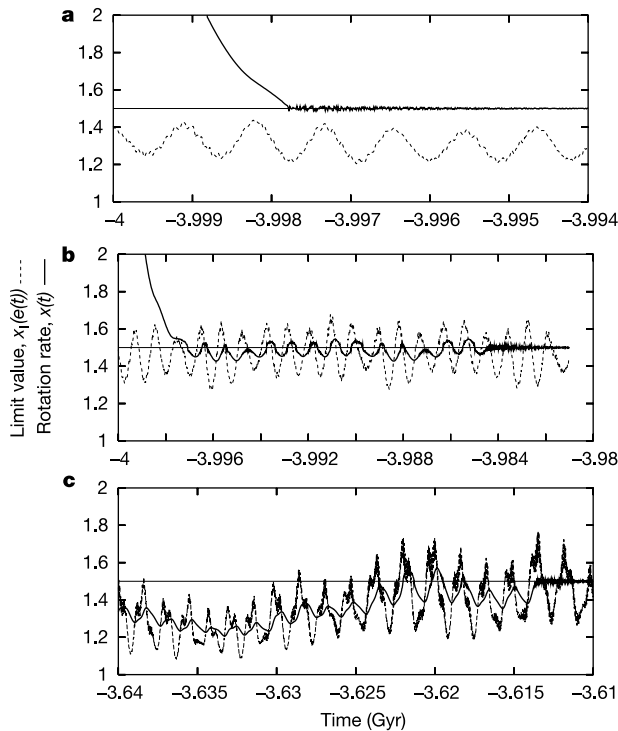


Figure 2 Typical cases of capture into the 3/2 resonance. The rotation rate $x(t)$ (bold curve) and limit value $x_l(e(t))$ (dotted curve) are plotted versus time (Gyr). **a**, Type I is the classical case⁵: As $e < e_{3/2}$, the limit value x_l is always lower than 3/2. **b**, In Type II, at the time when x reaches the resonant value 3/2, e is oscillating around $e_{3/2}$, leading to multiple crossings of the resonance, with ultimately a capture. **c**, Type III corresponds to solutions that have not been captured during the initial crossing of the resonance, but later on, as the eccentricity increases beyond $e_{3/2}$.

resonance p ; but later on, as the orbital elements evolve, the eccentricity increases beyond e_p , and tidal dissipation accelerates the spin rate beyond p , leading to additional capture (Fig. 2c).

Over 1,000 orbits, a few were initially trapped in high-order resonances (one in 7/2, one in 4/1, two in 9/2 and three in 5/1), but these were associated with high values of the planet's eccentricity. As the eccentricity decreased these resonances became unstable, and none of these high-order resonances survived. They did eventually get trapped for a long time into the 5/2 resonance, but even that did not survive over the full history of the planet. Indeed, the stability of the resonances depends on the eccentricity of Mercury, and except for the 1/1 resonance, the resonances may become unstable for very small values of the eccentricity (Table 1).

We followed all 1,000 solutions, starting from -4 Gyr, until they reached the present date or were captured into the 2/1, 3/2 or 1/1 resonances. Unlike in previous studies, we found that capture into the 1/1 resonance is possible, because the eccentricity of Mercury may decrease to very low values at which capture can occur and the resonance remain stable. Over 554 solutions that were captured into the 3/2 resonance, a single one, initially captured at -3.995 Gyr, escaped from resonance at about -2.396 Gyr. The solution then got trapped into the 1/1 resonance at -2.290 Gyr, capture that was favoured by the very low eccentricity required to destabilize the 3/2 resonance (Table 1). Out of the 56 solutions initially trapped into the 2/1 resonance, ten were destabilized, and only two of them were further captured, one into the 3/2 resonance, and one into the 1/1 resonance. Globally, we obtained a final capture probability:

$$P_{1/1} = 2.2\%, \quad P_{3/2} = 55.4\%, \quad P_{2/1} = 3.6\% \quad (1)$$

The remaining 38.8% non-resonant solutions end with nearly the same final rotation rate of $x_f = 1.21315$, because all the orbital

Table 1 Critical eccentricity $e_c(p)$ for the resonance p

p	$e_c(p)$
1/1	—
3/2	0.000026
2/1	0.004602
5/2	0.024877
3/1	0.057675
7/2	0.095959
4/1	0.135506
9/2	0.174269
5/1	0.211334

If $e < e_c(p)$, the resonance p becomes unstable, and the solution may escape the resonance. The critical eccentricity $e_c(p)$ is obtained by the resolution of $[\Omega(e)p - N(e)]/H(p, e) = (3/2K) \eta[(B - A)/C]$.

solutions are very close in the vicinity of the origin. Among the solutions captured into the 3/2 resonance, we can distinguish 31 solutions of type I, 168 of type II and 355 of type III (Fig. 2).

With the consideration of the chaotic evolution of the eccentricity of Mercury, we thus show that with a realistic tidal dissipative model that properly accounts for the damping of the libration of the planet, and without the need for some additional core–mantle friction, the present 3/2 resonant state is the most probable outcome for the planet.

Additionally, from the present state of the planet, we can derive an interesting constraint on its past evolution. Of all 554 orbits trapped into the 3/2 resonance, for 521 of them (94.0%) Mercury's eccentricity exceeded 0.325 in the past 4 Gyr. The conditional probability that Mercury's eccentricity exceeded 0.325, given that its rotation is trapped into the 3/2 resonance, is thus 94.0%. The 3/2 resonant state of Mercury thus becomes an observational clue that the chaotic evolution of the planet orbit led its eccentricity beyond 0.325 over its history.

The largest unknown in this study remains the dissipation factor k_2/Q of K (equation (4)) (ref. 15). A stronger dissipation would increase the probability of capture into the 3/2 resonance, because $x(t)$ would follow more closely $x_l(e(t))$ (Fig. 2), whereas lower dissipation would slightly decrease the capture probability. This study should apply more generally to any extrasolar planet or satellite whose eccentricity is forced by planetary perturbations. $\square$

Methods

Tidal dissipation and core–mantle friction will drive Mercury's obliquity (the angle between the equator and the orbital plane) close to zero. For zero-degree obliquity, and in the absence of dissipation, the averaged equation for the rotational motion near resonance p (where p is a half-integer) is^{4,5}:

$$\dot{x} = -\frac{3}{2}n \frac{B-A}{C} H(p, e) \sin 2(\ell - pM) \quad (2)$$

where ℓ is the rotational angle, $x = \dot{\ell}/n$ is the ratio of the rotation rate to the mean motion n , M is the mean anomaly and $H(p, e)$ are Hansen coefficients^{5,16}. The moments of inertia are $A < B < C$, with $C = \xi m R^2$, where m and R are the mass and radius of the planet, and ξ is a structure constant.

Tidal models independent of the frequency (constant- Q models) do not account for the damping of the amplitude of libration that is at present observed on Mercury^{5,17}. Moreover, these models introduce discontinuities into the equations and can thus be considered as unrealistic approximations for slow rotating bodies¹⁸. Therefore, we use here for slow rotations a viscous tidal model, with a linear dependence on the tidal frequency. Its contribution to the rotation rate is given by^{5,18,19,20}:

$$\dot{x} = -K[\Omega(e)x - N(e)] \quad (3)$$

with $\Omega(e) = (1 + 3e^2 + 3e^4/8)/(1 - e^2)^{9/2}$, $N(e) = (1 + 15e^2/2 + 45e^4/8 + 5e^6/16)/(1 - e^2)^6$, and

$$K = 3n \frac{k_2}{\xi Q} \left(\frac{R}{a} \right)^3 \left(\frac{m_0}{m} \right) \quad (4)$$

where k_2 and Q are the second Love number and the quality factor, while a , m and m_0 are the semi-major axis, the mass of the planet and the solar mass, respectively. Equilibrium is achieved when $\dot{x} = 0$, that is, for constant e , when $x = x_l(e) = N(e)/\Omega(e)$.

For a non-constant eccentricity $e(t)$, the limit solution of equation (3) is no longer $x_l(e)$, but more generally:

$$\tilde{x}_l(t) = \left(x(0) + K \int_0^t N(e(\tau)) g(\tau) d\tau \right) / g(t) \quad (5)$$

where $g(t) = \exp(K \int_0^t \Omega(e(\tau)) d\tau)$.

Using $\xi = 0.3333$, $k_2 = 0.4$ and $Q = 50$ (refs 15, 21), we have $K = 8.45324 \times 10^{-7} \text{ yr}^{-1}$. Assuming an initial rotation period of Mercury of 10 h, we estimated that the time needed to despin the planet to the slow rotations would be about 300 million years. This is why we started our integrations in the slow-rotation regime, with a rotation period of 20 days ($x \approx 4.4$) and a starting time of -4 Gyr , although these values are not critical.

Received 12 March; accepted 4 May 2004; doi:10.1038/nature02609.

- Pettengill, G. H. & Dyce, R. B. A radar determination of the rotation of the planet Mercury. *Nature* **206**, 1240 (1965).
- McGovern, W. E., Gross, S. H. & Rasool, S. I. Rotation period of the planet Mercury. *Nature* **208**, 375 (1965).
- Colombo, G. Rotation period of the planet Mercury. *Nature* **208**, 575 (1965).
- Colombo, G. & Shapiro, I. I. The rotation of the planet Mercury. *Astrophys. J.* **145**, 296–307 (1966).
- Goldreich, P. & Peale, S. J. Spin orbit coupling in the Solar System. *Astron. J.* **71**, 425–438 (1966).
- Counselman, C. C. & Shapiro, I. I. Spin-orbit resonance of Mercury. *Symp. Math.* **3**, 121–169 (1970).
- Goldreich, P. & Peale, S. J. Spin-orbit coupling in the solar system 2. The resonant rotation of Venus. *Astron. J.* **72**, 662–668 (1967).
- Peale, S. J. & Boss, A. P. A spin-orbit constraint on the viscosity of a Mercurian liquid core. *J. Geophys. Res.* **82**, 743–749 (1977).
- Anderson, J. D., Colombo, G., Espitio, P. B., Lau, E. L. & Trager, G. B. The mass, gravity field, and ephemeris of Mercury. *Icarus* **71**, 337–349 (1987).
- Brouwer, D. & Van Woerkom, A. J. J. The secular variations of the orbital elements of the principal planets. *Astron. Pap. Am. Ephem.* **XIII**, part II, 81–107 (1950).
- Bretagnon, P. Termes à longue périodes dans le système solaire. *Astron. Astrophys.* **30**, 141–154 (1974).
- Laskar, J. The chaotic motion of the solar system. *Icarus* **88**, 266–291 (1990).
- Laskar, J. Large-scale chaos in the Solar System. *Astron. Astrophys.* **287**, L9–L12 (1994).
- Laskar, J. *et al.* Long term evolution and chaotic diffusion of the insolation quantities of Mars. *Icarus* (in the press).
- Spohn, T., Sohl, F., Wiczerkowski, K. & Conzelmann, V. The interior structure of Mercury: what we know, what we expect from BepiColombo. *Planet. Space Sci.* **49**, 1561–1570 (2001).
- Hansen, P. A. Entwicklung der products einer potenz des radius vectors mit dem sinus oder cosinus eines vielfachen der wahren anomalie in reihen. *Abhandl. K. S. Ges. Wissensch.* **IV**, 182–281 (1855).
- Murray, C. D. & Dermott, S. F. *Solar System Dynamics* (Cambridge Univ. Press, Cambridge, 1999).
- Correia, A. C. M. & Laskar, J. Neron de Surgy, O. Long term evolution of the spin of Venus. I. Theory. *Icarus* **163**, 1–23 (2003).
- Munk, W. H. & MacDonald, G. J. F. *The Rotation of the Earth; A Geophysical Discussion* (Cambridge Univ. Press, Cambridge, 1960).
- Kaula, W. Tidal dissipation by solid friction and the resulting orbital evolution. *J. Geophys. Res.* **2**, 661–685 (1964).
- Yoder, C. F. Astrometric and geodetic properties of Earth and the Solar System. *Glob. Earth Physics: A Handbook of Physical Constants* 1–31 (American Geophysical Union, Washington DC, 1995).

Acknowledgements This work was supported by PNP-CNRS, Paris Observatory CS, and Fundação para a Ciência e a Tecnologia, POCTI/FNU, Portugal. The numerical computations were made at IDRIS-CNRS, and Paris Observatory. Authors are listed in alphabetic order.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be sent to J.L. (Laskar@imcce.fr).

Laser-induced ultrafast spin reorientation in the antiferromagnet TmFeO_3

A. V. Kimel¹, A. Kirilyuk¹, A. Tsvetkov¹, R. V. Pisarev² & Th. Rasing¹

¹NSRIM Institute, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

²Ioffe Physico-Technical Institute, 194021 St.-Petersburg, Russia

All magnetically ordered materials can be divided into two primary classes: ferromagnets^{1,2} and antiferromagnets³. Since ancient times, ferromagnetic materials have found vast application areas⁴, from the compass to computer storage and more recently to magnetic random access memory and spintronics⁵. In contrast, antiferromagnetic (AFM) materials, though representing the overwhelming majority of magnetically ordered materials, for a long time were of academic interest only. The fundamental difference between the two types of magnetic

materials manifests itself in their reaction to an external magnetic field—in an antiferromagnet, the exchange interaction leads to zero net magnetization. The related absence of a net angular momentum should result in orders of magnitude faster AFM spin dynamics^{6,7}. Here we show that, using a short laser pulse, the spins of the antiferromagnet TmFeO_3 can indeed be manipulated on a timescale of a few picoseconds, in contrast to the hundreds of picoseconds in a ferromagnet^{8–12}. Because the ultrafast dynamics of spins in antiferromagnets is a key issue for exchange-biased devices¹³, this finding can expand the now limited set of applications for AFM materials.

To deflect the magnetization of a ferromagnet from its equilibrium, a critical field $H_{\text{cr}}^{\text{FM}} \approx H_A$ of the order of the effective anisotropy field is required. In contrast, the response of an antiferromagnet to an applied field remains very weak before the exchanged-enhanced critical field $H_{\text{cr}}^{\text{AFM}} \approx \sqrt{H_A H_{\text{ex}}}$ is reached. In most materials the exchange field $H_{\text{ex}} \gg H_A$ ($H_A < 1 \text{ T}$, $H_{\text{ex}} \approx 100 \text{ T}$) and thus $H_{\text{cr}}^{\text{AFM}} \gg H_{\text{cr}}^{\text{FM}}$. This difference is related to the fact that in an antiferromagnet, no angular momentum is associated with the AFM moment. This large rigidity of an antiferromagnet to an external field also shows up in the magnetic resonance frequency¹⁴, where spin excitations start at $\omega \approx \gamma \sqrt{H_A H_{\text{ex}}}$. This is in contrast to $\omega \approx \gamma H_A$ in a ferromagnet, which can result in a difference of more than two orders of magnitude.

Indeed, dynamical many-body theory calculations¹⁵ show a possibility of AFM dynamics with a time constant of a few femtoseconds only. Experimentally, the ultrafast dynamics of an antiferromagnet is still an intriguing question. The problem however is far from trivial, as there is no straightforward method for the manipulation and detection of spins in AFM materials. Therefore, an appropriate mechanism should be found that would deflect the AFM moments on a timescale down to femtoseconds, and this change should subsequently be detected on the same timescale.

The solution to this problem can be found in the magnetocrystalline anisotropy. Indeed, a rapid change of this anisotropy can lead, via the spin–lattice interaction, to a reorientation of the spins^{11,12}. Such anisotropy change, in turn, can be induced by a short femtosecond laser pulse in a material with a strong temperature-dependent anisotropy. The subsequent reorientation of the spins can be detected with the help of time-resolved linear magnetic birefringence¹⁶, which enables us to follow the change of the direction of spins in antiferromagnets, similar to the Faraday and Kerr effects in ferromagnets.

The rare-earth orthoferrites RFeO_3 (where R indicates a rare-earth element) investigated here are known for a strong temperature-dependent anisotropy^{17,18}. These materials crystallize in an orthorhombically distorted perovskite structure, with a space-group symmetry D_{2h}^{16} ($Pbnm$). The iron moments order antiferromagnetically, as shown in Fig. 1, but with a small canting of the spins on different sublattices. The temperature-dependent anisotropy energy has the form^{19,20}:

$$\Phi(T) = \Phi_0 + K_2(T)\sin^2\theta + K_4\sin^4\theta \quad (1)$$

where θ is the angle in the x – z plane between the x axis and the AFM moment $\mathbf{G}$, see Fig. 1, and K_2 and K_4 are the anisotropy constants of second and fourth order, respectively. Applying equilibrium conditions to equation (1) yields three temperature regions corresponding to different spin orientations:

$$\Gamma_4(G_x, F_z): \theta = 0, T \geq T_2$$

$$\Gamma_2(G_z, F_x): \theta = 1/2\pi, T \leq T_1$$

$$\Gamma_{24}: \sin^2\theta = K_2(T)/2K_4, T_1 \leq T \leq T_2 \quad (2)$$

where T_1 and T_2 are determined by the conditions $K_2(T_1) = -2K_4$

and $K_2(T_2) = 0$ and the Γ s indicate the representations of the respective symmetry groups²¹. Therefore, depending on the anisotropy constants, a spin reorientation may be expected that shows two second-order phase transitions at T_1 and T_2 . The temperature dependence of $\sin^2\theta$ in the phase Γ_{24} is roughly linear²².

Of the possible resulting spin configurations compatible with the crystallographic symmetry, two are realized in several orthoferrites, in particular TmFeO_3 (shown in Fig. 1). The first, $\Gamma_4(G_xF_z)$, consists of a G -type AFM arrangement with the spins slightly canted from the $\pm x$ axes (resulting in a small net magnetic moment M_z) and is realized at temperatures above 91 K. In the second, $\Gamma_2(G_zF_x)$, that appears below 80 K, the spins lie nearly along the $\pm z$ axes. Temperature-dependent changes in the single-ion anisotropy favour the Γ_2 configuration at lower temperatures.

Experimentally, the transition between the two spin configurations in the antiferromagnet can be monitored with the help of linear optical birefringence. For light propagating along the z axis through a birefringent medium, the refractive indices for light polarized along the x axis and y axis are different, where Δn_{xy} characterizes the birefringence and is determined by a difference in the diagonal components of the dielectric permittivity tensor:

$$\Delta n_{xy} = (\varepsilon_{xx} - \varepsilon_{yy})/2n \quad (3)$$

Concerning a change of the diagonal components induced by the presence of an AFM vector $\mathbf{G}$, that is, $\varepsilon_{ii} = \varepsilon_{ii}^{(0)} + \beta_{ijk} G_j G_k$, one can find for TmFeO_3 that $\Delta n_{xy}(\Gamma_4) \neq \Delta n_{xy}(\Gamma_2)$. Thus, the birefringence serves as a direct measure of the orientation of the AFM vector $\mathbf{G}$.

At a photon energy of 1.55 eV, the pump laser pulse is absorbed in TmFeO_3 via the excitation of the localized electronic states of the Fe^{3+} and Tm^{3+} ions. The resulting changes of the birefringence are summarized in Fig. 2. In the time domain, the relaxation process can be divided into three distinct regions. First, the excitation decays via phonon cascades and the phonon system thermalizes in a very short time (process (1) in Fig. 2, with a time constant of around 0.3 ps; this time is in approximate agreement with earlier experiments²³). The phonon–phonon interaction sets a new lattice temperature so that the equilibrium anisotropy axis is changed.

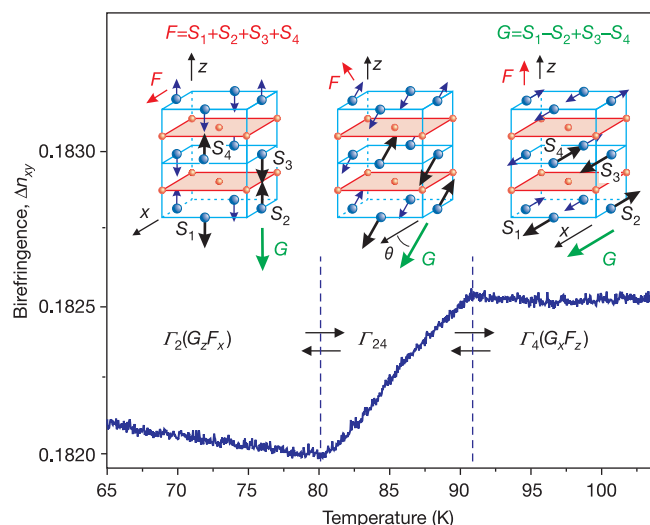


Figure 1 Linear optical birefringence in TmFeO_3 as a function of temperature. Insets show the corresponding arrangement of spins. The change of birefringence in the range of 80–91 K is due to the reorientation of the AFM vector in the x – z plane. It thus can be used for monitoring the ultrafast behaviour of spins in the pump–probe experiments.

Under such conditions in a ferromagnetic material, the magnetization vector would precess around its new equilibrium direction, approaching it because of the damping²⁴ (see Fig. 3a). In an antiferromagnet the resulting motion of the spins to the new equilibrium should occur in the plane spanned by $\mathbf{H}_A$ and $\mathbf{S}$ (see Fig. 3a). This is marked process (2) in the time dependencies of Fig. 2 and has a characteristic time of about 4 ps. This relaxation time is in agreement with a homogeneous precession period of AFM spins of 12 ps. The amplitude of this spin reorientation averaged through the probed volume is about 30 degrees in our experiment (see Fig. 4). This value is obtained with the help of the static birefringence data (Fig. 1) under the assumption of uniform heating. Owing to absorption in the sample the temperature profile created by the pump is not uniform. The pump-induced temperature increase as a function of penetration depth x can be expressed as $\Delta T(x) = T_m \times \exp(-kx)$, where k is about 700 cm^{-1} and T_m is about 35 K (ref. 18). The expected response of the sample on such photo-excitation was estimated via the integration through the probed volume and shown in Fig. 4 as a function of sample temperature. One can see that calculations and experiment are in very good agreement. Note that such thermal behaviour of the

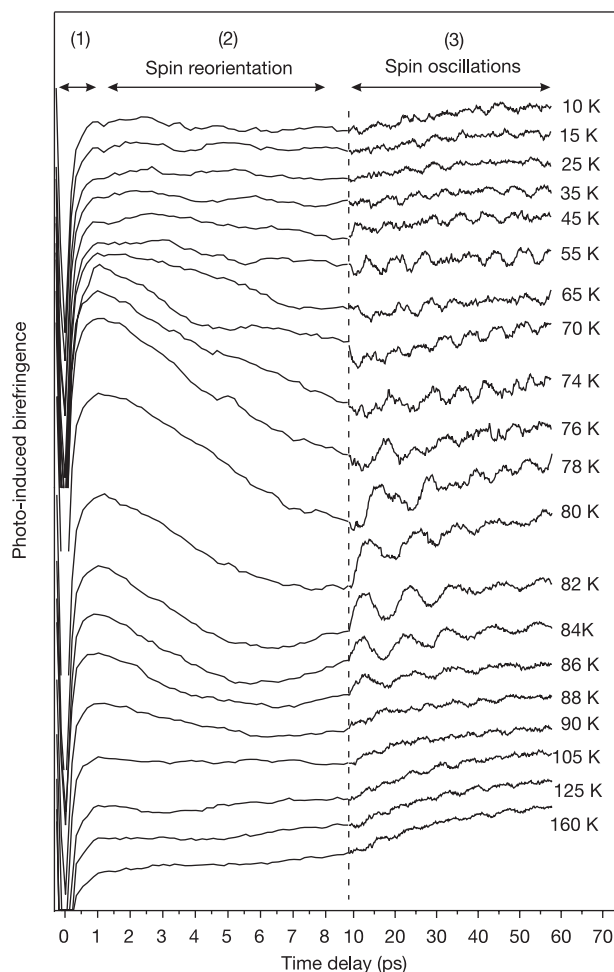


Figure 2 Excitation and relaxation of the AFM moment measured via changes in the magnetic birefringence. On the figure one can distinguish three processes: (1) electron–phonon thermalization with 0.3-ps relaxation time; (2) rotation of the AFM vector with 5-ps response time; (3) oscillations of the AFM vector around its equilibrium direction with an approximate 10-ps period.

response supposes that somewhere spins get fully reoriented over 90 degrees.

After the initial relaxation, we observe oscillations of the AFM moments around their new equilibrium (process (3)). Particularly strong oscillations are observable in the range of 80–90 K, that is, in the region of the re-orientational transition (Fig. 4). The strong temperature dependence of the frequency suggests that the oscillations have a magnetic origin and cannot be related to phonon modes. The temperature dependence of the derived frequencies resembles that of the spin waves with zero wave vector, shown by a thin line in Fig. 4 (ref. 20). Such spin waves are equivalent to a homogeneous precession of the magnetization observed at similar conditions in ferromagnets¹¹. It should be mentioned that the large amplitude of the oscillations observed and the inhomogeneous photo-excitation through the probed volume result in a difference between the theory and our experiment. The inset of Fig. 4 shows the calculated dependence of the AFM resonance frequency as a function of the initial amplitude²⁵. One can see that an increase of the amplitude results in a decrease of the frequency.

The experiment shows that the AFM spins in TmFeO_3 are reoriented by several tens of degrees within only a few picoseconds. For comparison, in a ferromagnet with an anisotropy energy similar to that in TmFeO_3 (10^4 J m^{-3} ; ref. 18), the magnetization precesses with a period of several hundred picoseconds^{8–10}. We should also note here that a relaxation time of about 700 ps was measured for the laser-induced destruction of the AFM order in FeBO_3 (ref. 26). In contrast, the spin reorientation happens to be a much faster process.

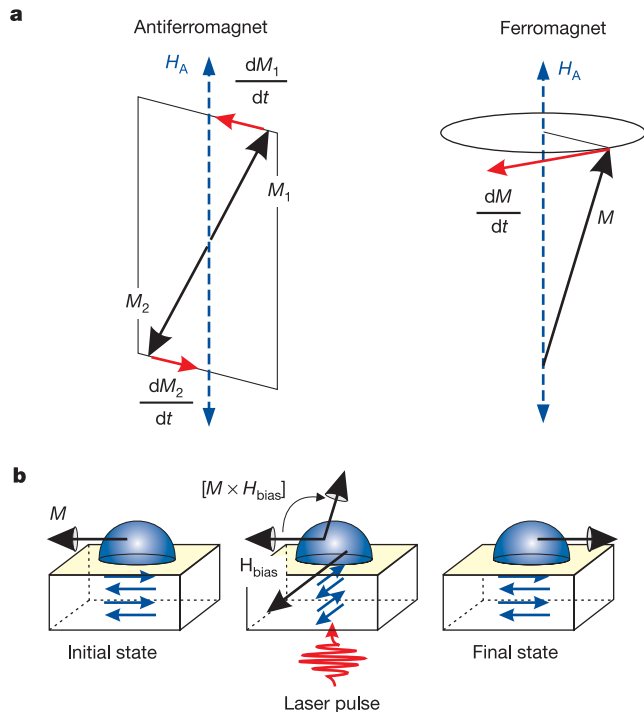


Figure 3 Precessional motion of spins in AFM and ferromagnetic compounds and its possible application. **a**, Schematic spin relaxation in an antiferromagnet compared to that in a ferromagnet: in contrast to the spiral ferromagnetic precession, the AFM vector moves in plane²⁷. M_1 and M_2 represent the magnetic moments of two AFM sublattices. Note that the Landau–Lifshitz equation has serious limitations in application to AFM resonance²⁵. **b**, Ultrafast laser-induced switching of a ferromagnetic/AFM bilayer: magnetization of the ferromagnetic layer precesses in the exchange field of the antiferromagnet into a final opposite state.

The measured maximum of the reorientation amplitude of 30 degrees, in fact, is only related to the problem of the instantaneous and homogeneous heating of a bulk sample and can easily be overcome in smaller structures, where such reorientation is of practical importance. For example, in an exchange-coupled ferromagnetic/AFM bilayer, the laser-induced reorientation of the AFM vector by 90 degrees will trigger, via the exchange coupling, a precession of the ferromagnetic moment into the opposite state¹². This possible application is illustrated in Fig. 3b. The reversal speed of the ferromagnet is determined by a precession of the ferromagnetic spins in the strong interfacial exchange field and does not require any other field. This principle could be integrated in magnetic random access memory when the heating is provided by a current pulse through the ferromagnet. Thus, in addition to increasing the stability of magnetic nanoelements¹³, the AFM layer could also play an active role in the ultrafast switching process of ferromagnetic nanostructures.

To conclude, we have demonstrated that an ultrafast spin reorientation in AFM TmFeO_3 can be induced by an ultrashort laser pulse. The rare-earth orthoferrites are an interesting and instructive class of materials with a strong temperature-dependent anisotropy, leading in some cases to a spin-reorientation transition. Optical excitation leads, via electron–phonon relaxation and phonon–phonon interaction, to a subpicosecond change of the anisotropy axis. Such change is equivalent to an ultrafast magnetic impact. We have shown the linear magnetic birefringence to be an excellent

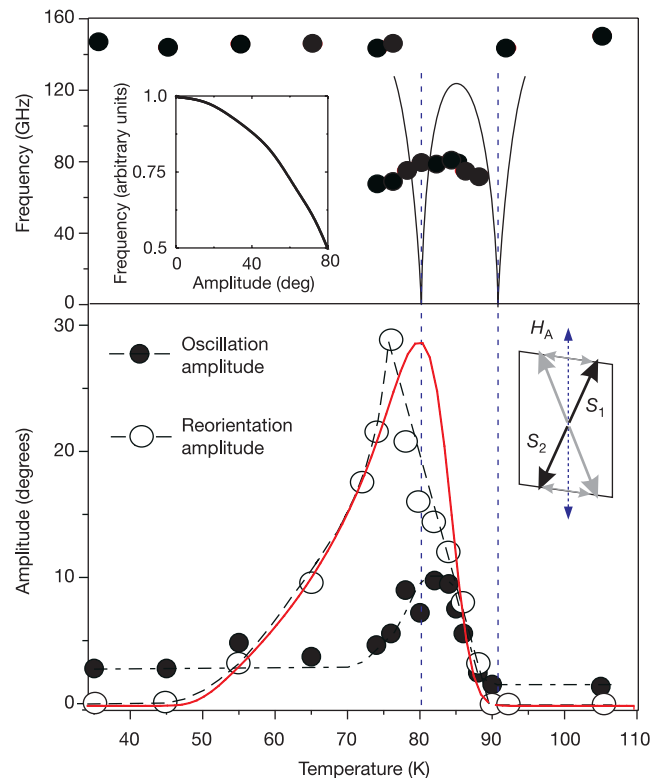


Figure 4 Temperature dependencies of the amplitudes and frequencies of the observed oscillations, as well as the amplitude of the spin reorientation. Black line shows the frequency change at the reorientational transition from ref. 19. Red line shows the expected response of the sample. The non-zero reorientation amplitude at $T = 55 \text{ K}$ corresponds to the instantaneous local laser-induced heating of about 35 K. Upper inset shows the calculated dependence of the AFM resonance frequency as a function of the initial amplitude²⁵. Lower inset shows the oscillations of spins in the x – z plane.

experimental technique to study the motion of the AFM vector. Thus the dynamics of the AFM moment can be influenced and detected with an all-optical pump–probe method. In contrast to a ferromagnet, the AFM spins can be fully reoriented within a few picoseconds' time. This observation of the ultrafast orientation dynamics of the AFM moment may have far-reaching consequences for future spintronic devices. □

Methods

The measurements were performed in a pump and probe configuration at a photon energy of 1.55 eV using amplified pulses from a Ti:sapphire laser at a repetition rate of 1 kHz. Each pulse had an energy of about 2 μ J and a profile close to gaussian, with a width at half-maximum of about 100 fs. The sample (a 60- μ m-thick TmFeO₃ single crystal cut perpendicular to the z axis) was placed in a cold finger cryostat where its temperature could be stabilized in the range of 10–300 K with a precision better than 0.5 K. The pump and probe beams were linearly polarized with an intensity ratio of about 100. Both beams were focused on the sample to a spot diameter of 50 μ m at half maximum for the pump and somewhat smaller for the probe beam. The probe, polarized at 45° with respect to the y axis in the sample plane, detected the birefringence changes induced by the pump, using a sensitive two-diode balanced detection scheme.

Received 30 December 2003; accepted 17 May 2004; doi:10.1038/nature02659.

- Weiss, P. L'hypothèse du champ moléculaire et la propriété ferromagnétique. (The hypothesis of the molecular field and the property of ferromagnetism.). *J. Phys. Rad.* **6**, 661–669 (1907).
- Stoner, E. C. *Magnetism and Matter* (Methuen, London, 1934).
- Néel, L. Propriétés magnétiques de l'état métallique et énergie d'interaction entre atomes magnétiques. *Ann. Phys. (5e Ser.)* **5**, 232–237 (1936).
- Jubilee issue. *J. Magn. Magn. Mater.* **200**, 1–789 (1999).
- Roukes, M. L. Electronics in a spin. *Nature* **411**, 747–748 (2001).
- Landau, L. D. & Lifshitz, E. M. *Electrodynamics of Continuous Media* (Pergamon, Oxford, 1960).
- Vonsovskii, S. V. *Magnetism* (John Wiley & Sons, New York, 1974).
- Gerrits, Th., van den Berg, H. A. M., Hohlfield, J., Bär, L. & Rasing, Th. Ultrafast precessional magnetization reversal by picosecond magnetic field pulse shaping. *Nature* **418**, 509–512 (2002).
- Schumacher, H. W. *et al.* Phase coherent precessional magnetization reversal in microscopic spin valve elements. *Phys. Rev. Lett.* **90**, 017201 (2003).
- Schumacher, H. W., Chappert, C., Sousa, R. C., Freitas, P. P. & Miltat, J. Quasiballistic magnetization reversal. *Phys. Rev. Lett.* **90**, 017204 (2003).
- van Kampen, M. *et al.* All-optical probe of coherent spin waves. *Phys. Rev. Lett.* **88**, 227201 (2002).
- Ju, G. *et al.* Ultrafast time resolved photoinduced magnetization rotation in a ferromagnetic/antiferromagnetic exchange coupled system. *Phys. Rev. Lett.* **82**, 3705–3708 (1999).
- Skumryev, V. *et al.* Beating the superparamagnetic limit with exchange bias. *Nature* **423**, 850–853 (2003).
- Kittel, C. Theory of antiferromagnetic resonance. *Phys. Rev.* **82**, 565 (1951).
- Ney, O., Trzeciak, M. & Hübner, W. Femtosecond dynamics of spin-dependent SHG response from NiO(001). *Appl. Phys. B* **74**, 741–744 (2002).
- Ferré, J. & Gehring, G. A. Linear optical birefringence of magnetic crystals. *Rep. Prog. Phys.* **47**, 513–611 (1984).
- White, R. L. Review of recent work on the magnetic and spectroscopic properties of the rare-earth orthoferrites. *J. Appl. Phys.* **40**, 1061–1069 (1969).
- Burzo, E. (ed.) *Numerical Data and Functional Relationships* Landolt-Börnstein New Series, Group III, Vol. 27f, 200–273 (Springer, Berlin, 1996).
- Horne, H. & Varma, C. M. Nature of spin-reorientation transitions. *Phys. Rev. Lett.* **20**, 845–846 (1968).
- Shapiro, S. M., Axe, J. D. & Remeika, J. P. Neutron-scattering studies of spin waves in rare-earth orthoferrites. *Phys. Rev. B* **10**, 2014–2021 (1974).
- Koster, G. F., Dimmock, J. O., Wheeler, R. G. & Statz, H. *Properties of the Thirty-Two Point Groups* (MIT Press, Cambridge, Massachusetts, 1963).
- Belov, K. P., Volkov, R. A., Goranskii, B. P., Kadamtseva, A. M. & Uskov, V. V. Nature of the transitions during the spontaneous reorientation of spins in rare-earth orthoferrites. *Fiz. Tverd. Tela* **11**, 1148–1151 (1969); *Sov. Phys. Solid State* **11**, 935–938 (1969).
- Kimel, A. V., Pisarev, R. V., Bentivegna, F. & Rasing, Th. Time-resolved nonlinear optical spectroscopy of Mn³⁺ ions in rare-earth hexagonal manganites RMnO₃ (R = Sc, Y, Er). *Phys. Rev. B* **64**, 201103 (2001).
- Landau, L. D. & Lifshitz, E. M. On the theory of the dispersion of magnetic permeability in ferromagnetic bodies. *Phys. Zeit. Sowjetunion* **8**, 153–169 (1935).
- Andreev, A. F. & Marchenko, V. I. Symmetry and the macroscopic dynamics of magnetic materials. *Usp. Fiz. Nauk.* **130**, 39–63 (1980); *Sov. Phys. Usp.* **23**, 21–34 (1980).
- Kimel, A. V., Pisarev, R. V., Hohlfield, J. & Rasing, Th. Ultrafast quenching of the antiferromagnetic order in FeBO₃: Direct optical probing of the phonon-magnon coupling. *Phys. Rev. Lett.* **89**, 287401 (2002).
- Yosida, K. *Theory of Magnetism* 125–143 (Springer, Berlin, 1998).

Acknowledgements This work was partially supported by the European IST network SPINOSA, the RTN network DYNAMICS, the Russian Foundation for Basic Research and Stichting voor Fundamenteel Onderzoek der Materie (FOM).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Th.R. (th.rasing@sci.kun.nl).

Reduction of hysteresis losses in the magnetic refrigerant Gd₅Ge₂Si₂ by the addition of iron

Virgil Provenzano*, Alexander J. Shapiro* & Robert D. Shull*

Magnetic Materials Group, NIST, 100 Bureau Drive, MS-8552, Gaithersburg, Maryland 20899, USA

* These authors contributed equally to this work

The magnetocaloric effect is the change in temperature of a material as a result of the alignment of its magnetic spins that occurs on exposure to an external magnetic field. The phenomenon forms the basis for magnetic refrigeration, a concept purported to be more efficient and environmentally friendly than conventional refrigeration systems^{1–5}. In 1997, a 'giant' magnetocaloric effect, between 270 K and 300 K, was reported in Gd₅Ge₂Si₂, demonstrating its potential as a near-room-temperature magnetic refrigerant^{6–8}. However, large hysteretic losses (which make magnetic refrigeration less efficient) occur in the same temperature range^{8,9}. Here we report the reduction (by more than 90 per cent) of these hysteretic losses by alloying the compound with a small amount of iron. This has the additional benefit of shifting the magnetic entropy change peak (a measure of the refrigerator's optimal operating temperature) from 275 K to 305 K, and broadening its width. Although the addition of iron does not significantly affect the refrigerant capacity of the material, a greater net capacity is obtained for the iron-containing alloy when the hysteresis losses are accounted for. The iron-containing alloy is thus a much-improved magnetic refrigerant for near-room-temperature applications.

The magnitude of the magnetocaloric effect is given by the field-induced entropy change, ΔS_m ; that is, when a magnetic field is applied to the material there is a decrease in its magnetic entropy due to the alignment of the spins with the field. The reduction in the magnetic entropy is compensated by an increase in the lattice entropy of the system (via the creation of phonons), resulting in a temperature increase. The reverse takes place upon the removal of the applied field. This excursion in temperature, ΔT , is the basis for magnetic refrigeration.

It is now fairly well accepted that the mechanism behind the large magnetocaloric effect in Gd₅Ge₂Si₂ involves a magnetic-field-induced crystallographic structural change from the high-temperature monoclinic paramagnetic phase to the low-temperature orthorhombic ferromagnetic phase^{8,10–12}. Because of the coincidence of sizeable hysteretic losses and large magnetocaloric effects in the same temperature range, it is reasonable to conclude that the same mechanism is responsible for both phenomena. Here we report a method for suppressing the crystallographic phase change and thereby greatly reducing the hysteresis losses in the Gd₅Ge₂Si₂ compound.

Figure 1 compares the X-ray powder diffraction patterns of Gd₅Ge₂Si₂ (trace I) and Gd₅Ge_{1.9}Si₂Fe_{0.1} (trace II) alloy samples that were homogenized in vacuum at 1,300 °C (1,573 K) for 1 h; on the top left of Fig. 1 are presented backscattered scanning electron microscopy (SEM) micrographs showing the typical microstructure of the two alloys. The X-ray diffraction patterns show that although very similar, the two spectra are slightly shifted along the abscissa with respect to each other, and most of the peaks of both materials match quite closely those of the monoclinic phase structure^{11–13}. The notable differences between the two patterns are the increased magnitude in the iron-containing alloy of the peak centred around 35° and the appearance of some smaller peaks (centred at about 30°,

31.6° and 35.6°) in the $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ alloy which are not present in the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ spectrum. These differences in the two X-ray spectra are due to the microstructural differences in the two alloys (see discussion below). In fact, as it can be seen from the SEM micrographs, the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound is observed to be single-phase (Fig. 1, left inset), whereas the microstructure of the alloy with the iron addition (Fig. 1, middle inset) is characterized by a dominant matrix (light grey) phase and by a minor (darker) phase located along the grain boundaries of the matrix phase. Closer examination of the smaller grain boundary phase in the iron-containing alloy shows it to be actually made up of two phases: a dark grey phase and a black phase (Fig. 1, right inset).

Energy-dispersive spectroscopy (EDS) analysis conducted on the two alloys yielded the following results. The composition, in atomic fraction, of the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound was found to be 55% Gd, 22.6% Ge and 22.4% Si; these compositional values are very close to those obtained on the same alloy by a wet chemistry method, and are also close to the target composition of 55.6% Gd, 22.2% Ge and 22.2% Si. By contrast, the compositional values of the three phases present in the compound containing iron are as follows: (1) for the dominant matrix phase — 60% Gd, 22.2% Ge, 17.8% Si and no iron; (2) for the lighter grain boundary phase — 35.3% Gd, 11.2% Ge, 25.6% Si and 27.9% Fe; (3) for the darker grain boundary phase — 22.2% Gd, 7.2% Ge, 34.4% Si and 36.2% Fe. Note that the dominant phase in the iron-containing compound had a larger content of both Gd and Ge and a smaller concentration of Si relative to the target composition of 55.6% Gd, 21.1% Ge, 22.2% Si and 1.1% Fe. On the other hand, the grain boundary phases were both rich in Si and Fe. These results strongly suggest that one of the major effects of the iron addition is for it to combine primarily with the silicon (giving rise to the two boundary phases), resulting in a matrix phase that is depleted in silicon compared to the single phase of the alloy without iron. The greater concentration of Gd, together with the

deficiency of Si in the matrix phase, probably plays a pivotal role in the different magnetic behaviour of the $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ alloy, compared to that of the compound without iron.

The two sets of hysteresis loops presented in Fig. 2 illustrate the hysteresis losses for the two alloys in the 260–320 K temperature range. In this temperature interval the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound exhibits large hysteretic losses (Fig. 2a), coinciding with the same temperature interval in which the corresponding magnetic entropy change, ΔS_m , attains its peak values. A summary of hysteretic loss values over the temperature range of interest is presented in Fig. 3. These hysteretic losses were determined by computing the hatched area inside each magnetization (M) versus field (H) loop shown in Fig. 2. It can be seen clearly that the addition of about one atomic per cent of iron to the alloy resulted in a reduction of the hysteresis losses by nearly 95 per cent when compared to the alloy without the iron addition.

Closer examination of the M – H loops shown in Fig. 2 provides additional insight concerning the effect of the iron addition on the magnetocaloric response of the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound in the 270–320 K temperature range. For the alloy without the iron addition (Fig. 2a), the magnetization measured in the first half of the loop shows a distinct magnetic transition with increasing field between 270 K and 290 K; this transition occurs at higher field values with increasing temperature. As already stated, it has been hypothesized that this transition is the result of a field-induced first-order crystallographic phase change from the paramagnetic monoclinic phase to a ferromagnetic orthorhombic phase^{8,10–12}. The M – H loops clearly show that this field-induced phase transition reverses upon decreasing the field, but with some hysteresis in the transition field. Below 270 K this alloy exhibits ferromagnetic behaviour, whereas above 295 K the material is paramagnetic.

By contrast, the M – H loops of the $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ alloy (Fig. 2b) do not show any field-induced magnetic transition in the 260–340 K

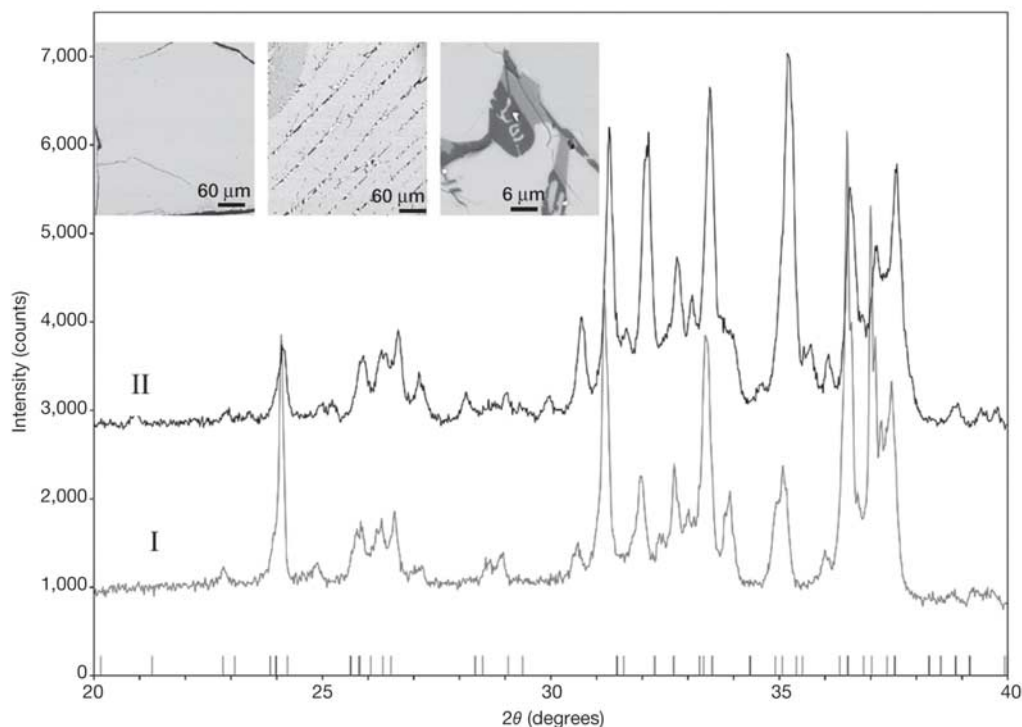


Figure 1 X-ray diffraction spectra and SEM micrographs of the two alloys. Cu-K α X-ray diffraction spectra measured at room temperature for the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound (I) and the $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ alloy (II). The Bragg peak positions for the monoclinic phase are shown along the x-axis. Insets, backscattered SEM micrographs, showing typical microstructures

of the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ (left inset) and $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ (middle and right inset) alloys; $\text{Gd}_5\text{Ge}_2\text{Si}_2$ is single phase, whereas the $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ alloy consists of a dominant grey phase and a minor, darker intergranular phase.

temperature range. In fact, here the magnetic data show a gradual shift from ferromagnetic behaviour to superparamagnetic behaviour, as evidenced by the appearance of curvature in the M - H plots with increasing T ; above 320 K, the material is paramagnetic. By comparison, the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound exhibits paramagnetic behaviour at temperatures above 290 K. Although not shown here, the M versus T data of the $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ alloy at $H = 796 \text{ kA m}^{-1}$ (1 T) exhibit no magnetic transition for temperatures below 260 K. The X-ray diffraction data (Fig. 1) and the magnetization data (Fig. 2) strongly suggest that the main effect of the iron addition in the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound is to suppress the field-induced monoclinic-to-orthorhombic phase transition in the 270–320 K range, resulting in much smaller hysteresis losses. This field-induced phase transition is absent in the iron-containing alloy because the orthorhombic phase never forms for $H < 3,980 \text{ kA m}^{-1}$ (5 T) between 10 K and 360 K, whereas it does form in $\text{Gd}_5\text{Ge}_2\text{Si}_2$ at $T < 270 \text{ K}$ without the presence of a magnetic field.

Variation of the magnetic entropy change, ΔS_m , with temperature for the two alloys is shown in Fig. 4. These data were computed from the isothermal M - H data of the two alloys using equation (3). For the alloy without iron, the value of the ΔS_m peak, integrated over an applied field $\Delta H = 3,980 \text{ kA m}^{-1}$ (5 T), is about a factor of 3

higher than that for the alloy with the iron ($20 \text{ J kg}^{-1} \text{ K}^{-1}$ versus $7 \text{ J kg}^{-1} \text{ K}^{-1}$, respectively). However, the ΔS_m peak width of the iron-containing alloy is considerably broader. For this latter alloy, the peak of ΔS_m occurs at about 305 K (Fig. 4b), whereas in the alloy without iron, the ΔS_m peak occurs nears 275 K (Fig. 4a). From the data presented in Fig. 4, the refrigerant capacity (RC) value was computed for the two alloys for $\Delta H = 3,980 \text{ kA m}^{-1}$ (5 T), using the two approaches described in the Methods section. The RC values computed by the expression given in ref. 14 were respectively around 360 J kg^{-1} and 305 J kg^{-1} for the alloys with and without the iron addition, whereas the RC values computed by the Wood and Potter method¹⁵ were respectively around 240 J kg^{-1} and 265 J kg^{-1} .

In the temperature ranges where the RC values were computed by both methods, the average hysteretic loss is about 65 J kg^{-1} for the compound without iron but less than 4 J kg^{-1} for the compound with the iron addition. Because these are the costs in energy to make one cycle of the magnetic field, they must be considered when calculating the usefulness of a magnetic refrigerant being subjected to cyclic fields. One way to take into account the hysteresis loss of each alloy is to simply subtract it from the corresponding RC value. Subtraction of the average hysteresis for each alloy from the corresponding RC values computed by the methods of refs 14 and 15 respectively yielded approximate values of 355 J kg^{-1} and 235 J kg^{-1} for the Fe-containing alloy, and approximate values of 240 J kg^{-1} and 200 J kg^{-1} for the alloy without the Fe addition. Note how much larger are the resultant values for the Fe-containing alloy regardless of the RC calculation method. Therefore, on this basis the $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ alloy is a much better magnetic refrigerant than the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound despite its lower ΔS_m peak value.

In order to examine further the effect of iron concentration on the magnetic properties of the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound, especially with regard to the hysteretic losses, two additional alloy samples were prepared. These alloys had respectively one-half and twice the iron concentration of $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$; they were arc-melted and homogenized as described in the Methods section for the other two alloys. Owing to space limitations, details of their microstructure and magnetic data are not presented here, but their magnetic properties can be summarized as follows. For the alloy having half the iron concentration, the M - H loops were similar to those of the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound—that is, they showed the presence of both large hysteresis losses and a field-induced magnetic phase transition in the 260–320 K temperature range. However, compared to the iron-free compound, the hysteresis losses were only about 50% smaller and the field-induced phase transition was less pronounced. The M - H loops of the alloy having twice the iron concentration had

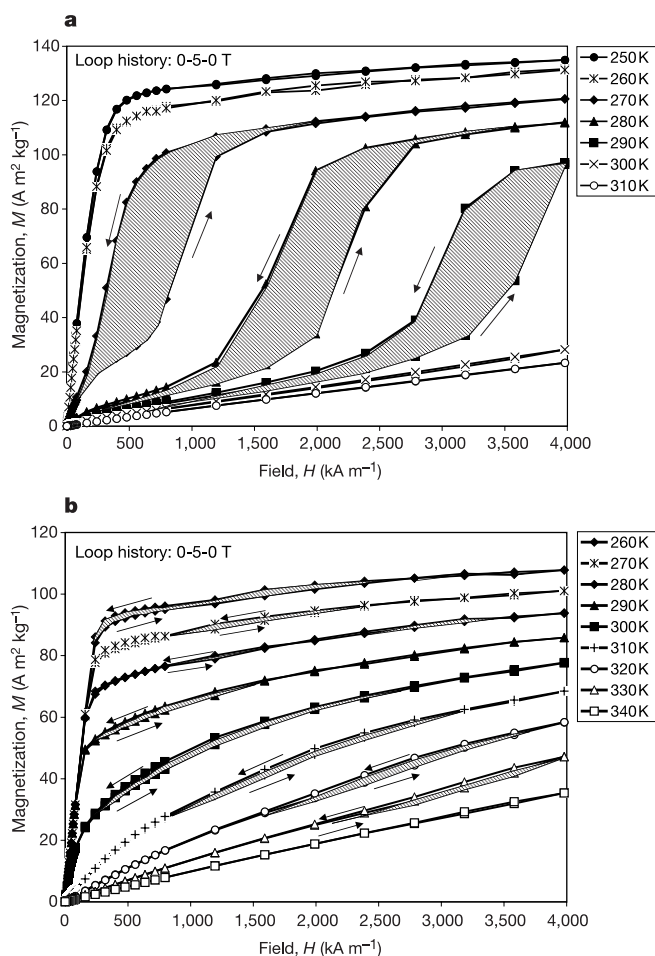


Figure 2 Magnetization versus field curves. Magnetization versus field curves for the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound between 250 K and 310 K (a) and for the $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ alloy between 260 K and 340 K (b); arrows indicate the sequence of measurements. The curves qualitatively illustrate the large hysteresis losses of the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound and the much smaller values of the $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ alloy. In addition, paramagnetic behaviour is observed above the different temperatures 320 K and 290 K for the Fe-containing alloy and $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound, respectively.

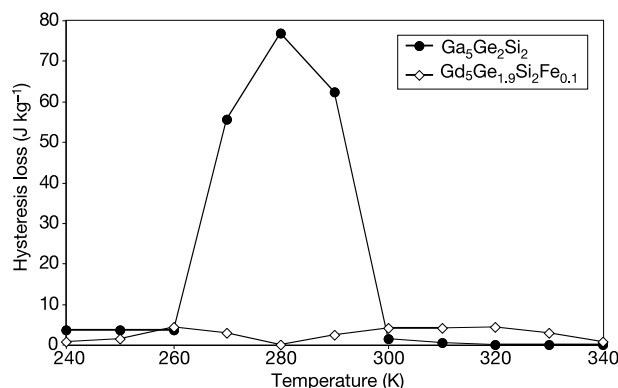


Figure 3 Comparison of hysteresis losses. Comparison of hysteresis losses (calculated as described in the text) of the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ (filled circles) and $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ (open diamonds) alloys plotted as a function of temperature.

almost double the hysteresis losses of the $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ alloy, and there was a slight indication of a field-induced phase transition in the 260–320 K temperature range. These results show a systematic variation of magnetic character with iron concentration, and show that the lowest hysteresis losses were observed for the alloy whose iron concentration was about 1 per cent atomic fraction.

As mentioned earlier, one of the major effects of the iron addition to the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound is to suppress the formation of the orthorhombic phase. In the $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ alloy, the monoclinic phase is the stable structure even at low temperatures, and the amount of orthorhombic phase that forms is quite small. In $\text{Gd}_5\text{Ge}_2\text{Si}_2$ the stable structure is the orthorhombic phase below 270 K and the monoclinic phase above 270 K. Therefore, as the amount of orthorhombic phase is very small in the iron-containing alloy, the monoclinic-to-orthorhombic field-induced phase transformation is not observed between 260 K and 340 K for H application up to $3,980 \text{ kA m}^{-1}$ (5 T), resulting in negligible hysteretic losses. The reverse is true for the alloy without the iron addition. This change in thermodynamics is probably related to the reduction in Si content and the increase in Gd content of the matrix phase when iron is added to the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound. The multiple-phase nature of the quaternary alloy probably also contributes to the stabilization of the monoclinic structure, perhaps by mediating the compressive stress created during cooling as the alloy contracts.

The Fe addition also creates a magnetic nanostructure. The

superparamagnetic behaviour observed in the $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ quaternary alloy (and not in the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound) at higher temperatures shows that the ferromagnetic matrix phase has been broken up into nanometre-sized ferromagnetic clusters. This type of magnetic nanocomposite morphology has been shown to lead to enhanced magnetocaloric effects^{4,16,17}, and is probably also the reason why the ΔS_m peak width of the iron-containing alloy is considerably broader. How such a magnetic structure developed is unknown. Certainly the microstructure does not show this nanoscale structure, so the cause must be more subtle (perhaps the inhomogeneous distribution of vacancies or defects in the Gd-Si-Ge matrix phase). □

Methods

The samples used in this study were prepared by arc melting, using a water-cooled copper hearth in an argon atmosphere under ambient pressure starting with the appropriate amounts of the component elements. The compound without the iron addition was prepared by the Materials Preparation Centre of the Ames Laboratory, Ames, Iowa, whereas the alloy with the iron addition was prepared at the alloy melting facility of the American Dental Association at NIST. The purity of the starting constituents was 99.9% mass fraction or better for the two alloys; and their target compositions were as follows: $\text{Gd}_5\text{Ge}_2\text{Si}_2$ and $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ (approximately one atomic per cent Fe). Before making magnetic measurements, the alloy samples were homogenized at $1,300^\circ\text{C}$ for one hour in vacuum⁷. Following the homogenizing treatment, the crystal structure and the microstructure of the samples were respectively characterized by Cu-K α powder X-ray diffraction (XRD) analysis, and by SEM and EDS. The magnetocaloric effect of the two alloys were determined by measuring M as a function of T and H , using a SQUID magnetometer. The magnetometer, calibrated by a pure Ni sphere SRM no. 772 from NIST, was programmed to measure M at discrete magnetic field values between 0 and $3,980 \text{ kA m}^{-1}$ (5 T) while the temperature was held constant. Then the measurement sequence was repeated many times successively at each 10 K interval as the temperature was raised from 10 K up to 360 K. The M values were normalized by dividing them by the corresponding sample mass. ΔS_m was then calculated from the M data using the integrated Maxwell relation:

$$\left(\frac{\partial S}{\partial H}\right)_T = \left(\frac{\partial M}{\partial T}\right)_H \quad (1)$$

Integration of the above expression leads to:

$$\Delta S_m(T, \Delta S) = \int_0^{H'} \left(\frac{\partial M}{\partial T}\right)_H dH \quad (2)$$

For M measurements made at constant temperature at discrete H intervals, the above Maxwell expression can be approximated by the following expression:

$$\Delta S_m \approx \frac{1}{\Delta T} \left[\int_0^{H'} M(T + \Delta T, H) dH - \int_0^{H'} M(T, H) dH \right] \quad (3)$$

Therefore, using the approximation given by equation (3), ΔS_m as a function of temperature for each sample was computed numerically by first differentiating the magnetization data, M , with respect to temperature, and then integrating the resulting derivatives from zero field to some value H up to 5 T.

We now discuss the experimental uncertainty in the measurements of H , T and M that directly affect the overall accuracy of the ΔS_m values computed by equation (3). Values of H between zero and $3,980 \text{ kA m}^{-1}$ (5 T) are known to within 0.2%. During each set of isothermal magnetic measurements where M was measured as a function of magnetic field, the temperature was kept constant to within 0.02 K. Accuracy of the magnetization data varied from 0.5% to 2% for both alloys, with most data having an uncertainty of less than 1% at non-zero field values. At zero field, the uncertainty ranged from 1% to 8%. Consequently, the error in ΔS_m is about 1% or less. We represent these errors by the sizes of the symbols in Figs 2–4. In Fig. 4b, below 250 K there is some oscillation in the ΔS_m versus T plot that is not a reflection of uncertainty in the data, but rather reflects real variations in the magnetization data, probably due to the secondary phases in the alloy. In fact, we prepared two alloys having compositions of the secondary phases detected in the Fe-containing alloy discussed here, and observed peaks in ΔS_m that correlate with the temperature regions of oscillation shown in Fig. 4b. These results will be presented elsewhere. Even though the data shown in Fig. 4 were calculated using measured M versus H data taken at 10 K intervals for comparison with earlier published data^{7,8}, the temperature of maximum ΔS_m was also computed from magnetic data measured at 4 K intervals, leading us to conclude that the temperature at which ΔS_m is maximum is accurate to within $\pm 5 \text{ K}$.

RC values for the two alloys were determined by two different methods. In the first method¹⁴, the RC values were obtained by numerically integrating the area (shaded in Fig. 4) under the ΔS_m versus T curves, using the temperatures at half-maximum of the ΔS_m peak as the integration limits. In the second method, the RC values were computed using the approach suggested by Wood and Potter¹⁵ and later used by other researchers in the field^{18,19}. Wood and Potter defined the refrigerant capacity for a reversible refrigeration cycle operating between T_h (the temperature of the hot reservoir) and T_c (the temperature of the cold reservoir) as $\text{RC} = \Delta S_m \Delta T$, where ΔS_m is the magnetic entropy change at the hot and cold ends of the cycle (defined equal) and $\Delta T = T_h - T_c$. According to this approach, for a given magnetic refrigerant the optimum refrigeration cycle occurs when

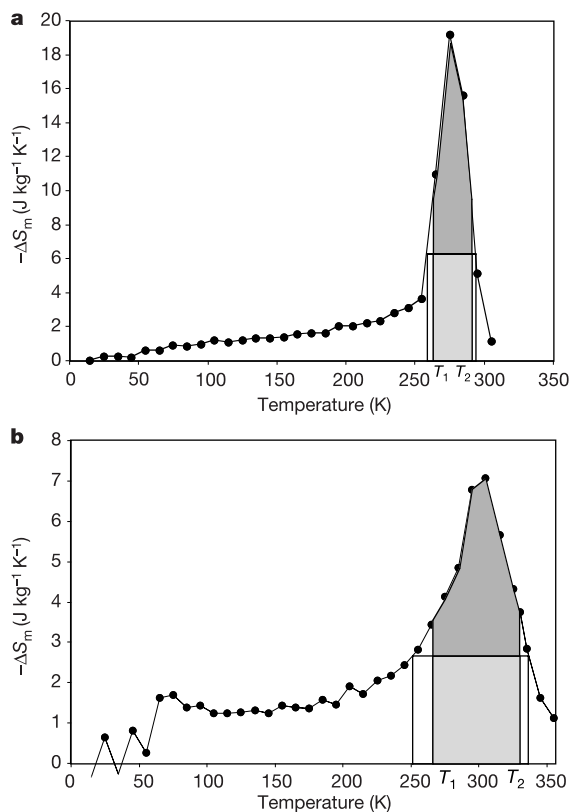


Figure 4 Computed ΔS_m and RC values. Computed ΔS_m (for $\Delta H = 3,980 \text{ kA m}^{-1}$ (5 T)), normalized with respect to sample mass, of the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound (**a**) and of the $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ alloy (**b**), plotted as a function of temperature. Note the presence of peaks centred near 270 K and 305 K respectively for the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound and the $\text{Gd}_5\text{Ge}_{1.9}\text{Si}_2\text{Fe}_{0.1}$ alloy. The refrigerant capacities (RC), calculated as described in the Methods section, are also shown as the shaded areas and unshaded rectangles for each material; T_1 and T_2 refer to the limits of integration for calculating the shaded areas. Compared to the $\text{Gd}_5\text{Ge}_2\text{Si}_2$ compound, the iron-containing compound exhibits a broader ΔS_m peak and a ΔS_m peak at a higher temperature.

the quantity $\Delta S_m \Delta T$ is maximum. The RC values computed by the Wood and Potter method are shown in Fig. 4 as the rectangular areas overlapping and extending outside the shaded areas.

Received 2 December 2003; accepted 5 December 2004; doi:10.1038/nature02657.

- Glanz, J. Making a bigger chill with magnets. *Science* **279**, 2045 (1998).
- Tegus, O., Bruck, E., Buschow, K. H. J. & de Boer, F. R. Transition-metal-based refrigerants for room-temperature applications. *Nature* **415**, 150–152 (2002).
- Tegus, O. *et al.* Magnetic-phase transitions and magnetocaloric effects. *Physica B* **319**, 174–192 (2002).
- McMichael, R. D., Shull, R. D., Swartzendruber, L. J., Bennett, L. H. & Watson, R. E. Magnetocaloric effect in superparamagnets. *J. Magn. Magn. Mater* **111**, 29–33 (1992).
- Fujita, A., Fujieda, S., Hasegawa, Y. & Fukamichi, K. Itinerant-electron metamagnetic transition and large magnetocaloric effects. *Phys. Rev. B* **67**, 104416 (2003).
- Pecharsky, V. K. & Gschneider, K. A. Jr The giant magnetocaloric effect in $Gd_5(Si_2Ge_2)$. *Phys. Rev. Lett.* **78**, 4494–4497 (1997).
- Perchinsky, A. O., Gschneider, K. A. Jr & Pecharsky, V. K. The giant magnetocaloric effect of optimally prepared $G_5Ge_2Si_2$. *J. Appl. Phys.* **93**, 4722–4728 (2003).
- Levin, E. M., Pecharsky, V. K. & Gschneider, K. A. Jr Unusual magnetic behavior in $G_5(Ge_{1.5}Si_{2.5})$ and $G_5Ge_2Si_2$. *Phys. Rev. B* **62**, R14625–R14628 (2000).
- Giguere, A. *et al.* Direct measurements of the 'giant' adiabatic temperature change in $G_5Ge_2Si_2$. *Phys. Rev. Lett.* **83**, 2262–2265 (1999).
- Choe, W. *et al.* Making and breaking covalent bonds across the magnetic transition in the giant magnetocaloric material $Gd_5(Ge_2Si_2)$. *Phys. Rev. Lett.* **84**, 4617–4620 (2000).
- Morellon, L., Algarabel, P. A., Ibarra, M. R., Blasco, J. & Garcia-Landa, B. Magnetic-field-induced structural transition in $Gd_5(Si_{1.8}Ge_{2.2})$. *Phys. Rev. B* **58**, R14721–R14724 (1998).
- Meyers, J., Chumbley, S., Choe, W. & Miller, G. J. Microstructural analysis of twinned β - $G_5Ge_2Si_2$. *Phys. Rev. B* **66**, 012106 (2002).
- Pecharsky, V. K. & Gschneider, K. A. Jr Phase relationship and crystallography in pseudobinary system Gd_5Si_4 - Gd_5Ge_4 . *J. Alloys Compd.* **260**, 98–106 (1997).
- Gschneider, K. A. Jr, Pecharsky, V. K., Pecharsky, A. O. & Zimm, C. B. Recent development in magnetic refrigeration. *Mater. Sci. Forum* **315–317**, 69–76 (1999).
- Wood, M. E. & Potter, W. H. General analysis of magnetic refrigeration and its optimization using a new concept: maximization of refrigerant capacity. *Cryogenics* **25**, 667–683 (1985).
- Shull, R. D., Swartzendruber, L. J. & Bennett, L. H. in *Proc. Sixth Int. Cryocoolers Conf.* (eds Green, G. & Knox, M.) 231 (Publication no. DTRC-91/002, David Taylor Research Center, Annapolis, Maryland, 1991).
- McMichael, R. D., Ritter, J. J. & Shull, R. D. Enhanced magnetocaloric effect in $Gd_5Ga_{5-x}Fe_xO_{12}$. *J. Appl. Phys.* **73**, 6946–6948 (1993).
- Tishin, A. M. Magnetic refrigeration in the low-temperature range. *J. Appl. Phys.* **68**, 6480–6484 (1990).
- von Ranke, P. J. *et al.* Anomalous magnetocaloric effect in YbAs associated with the giant quadrupolar interaction. *Phys. Rev. B* **63**, 24422 (2001).

Acknowledgements This work was partially supported by the NASA-Goddard Space Flight Center. We thank R. Waterstrat and R. Drew for technical assistance, and M. Lufaso and R. D. McMichael for helpful discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to V.P. (virgil12@nist.gov).

Evidence for a macroscopic electric field in the sedimentation profiles of charged colloids

Mircea Raşa & Albert P. Philipse

Van 't Hoff Laboratory for Physical and Colloid Chemistry, Debye Institute, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

The determination of molecular masses from barometric sedimentation profiles, a main topic in ultracentrifugal analysis, is thought to be quantitatively correct for non-interacting particles^{1,2}. Whereas this expectation is justified for uncharged colloids or macromolecules at low volume fractions, early ultracentrifugation studies³ on charged particles had already indicated that the obtained masses might be much too low. More recently, expanded sedimentation profiles have been observed for charged particles^{4,5}, sometimes inflated by orders of magnitude⁵ relative to the barometric prediction, which highlights a short-

coming in our understanding of centrifugation of even very dilute charged species⁵. Theory⁶ and simulations⁷, anticipated by various authors^{4,8,9}, now propose that strongly non-barometric sedimentation profiles might be caused by an internal macroscopic electric field that, even for non-interacting particles, significantly decreases the buoyant particle mass. The existence of this field and its intriguing consequences still lack experimental verification. Here we report ultracentrifugation experiments on charged colloidal silica spheres, showing both the existence of such a macroscopic electric field and its drastic effects on the sedimentation profiles of very dilute dispersions at low ionic strength.

Centrifugation is an indispensable technique for separating and analysing cells, organelles and macromolecules^{1,2}, as well as colloids¹⁰. A classical example is the centrifugation of DNA fragments in salt gradients, which confirmed the Watson–Crick model for DNA replication^{11,12}. Absolute molecular masses can be determined directly under non-denaturing conditions from sedimentation–diffusion (SD) concentration profiles, a method reported to be rigorous². The method assumes that, for sedimentation under gravity, the particle number density $\rho(x)$ at an altitude x in the SD profile follows from a Boltzmann distribution and has the form

$$\ln(\rho(x)) \propto -x/L \quad (1)$$

In this 'barometric' profile, $L = k_B T / (mg)$ is the gravitational length for particles with buoyant mass m , T is the absolute temperature, k_B is the Boltzmann constant and g is the gravitational acceleration. It is generally assumed¹⁰ that non-interacting particles will adopt such a barometric profile and that, consequently, sufficiently low concentrations ensure the validity of equation (1). Here, however, we report SD profiles of very dilute charged colloids that strongly deviate from the barometric distribution (equation (1)) owing to an electric field, which has only recently^{6,7,13} been clearly identified as an important factor in the centrifugation of charged species.

Non-barometric behaviour due to an electric field is predicted to occur at sufficiently low ionic strength⁶, and therefore we have studied the ultracentrifugation of well-defined, charged silica spheres (Table 1) in ethanol. This solvent is suitable because of its inherent low ionic strength and its ability to disperse charged silica spheres to non-aggregated 'alcosols' with practically unlimited colloidal stability^{5,14,15}. We found that the centrifuged silica spheres form reproducible SD profiles that can be scanned with high spatial resolution (Fig. 1; see Methods section). SD profiles were studied for initial silica volume fractions down to 0.01% to minimize the effect of inter-particle interactions. We verified that the barometric part (region I; see below) of SD profiles yields a correct colloid radius: for a silica mass density of 1.6 g cm^{-3} (Table 1), the measured centrifugal lengths correspond to a radius of 19.2 nm, which lies within the radius distribution determined from electron microscopy (Table 1). We also verified that uncharged silica spheres dispersed in cyclohexane yield the expected barometric profiles. However, for charged silica spheres all our experimental SD profiles, with a representative selection in Fig. 1, deviate drastically from the barometric distribution.

These deviations are due to a macroscopic electric field in the SD profile, that is, a gradient in an equilibrium electrical potential,

Table 1 Properties of silica spheres (Labcode SiA) dispersed in ethanol

R (nm)	σ (%)	R_h (nm)	δ (g cm^{-3})	μ ($\mu\text{m cm V}^{-1} \text{s}^{-1}$)	z
21.9	11.6	30.0	1.6 ± 0.1	-0.95 ± 0.05	≈ 50

R and σ are radius and polydispersity, respectively, from transmission electron microscopy; R_h is the radius from dynamic light scattering; δ is the mass density from ref. 14; μ is the electrophoretic mobility for silica volume fractions in the range 0.01–0.3%; z is the number of elementary charges on silica particles, determined from electrophoretic mobility. Errors are $\pm$ s.d.

unrelated to the more familiar non-equilibrium streaming potentials or sedimentation potentials¹⁰ generated by solvent motion relative to the colloids. The equilibrium field can be understood as a consequence of the huge mass difference between the (in our case negative) colloids and their positive counterions⁶. The practically weightless counterions attempt to achieve a homogeneous distribution in a sedimentation cell to maximize their entropy, whereas the heavy colloids accumulate near the cell bottom to minimize their gravitational potential energy. Electroneutrality, as has been pointed out⁶, couples the two tendencies, causing an 'entropic lift' for the colloids, that is, an effective increase in their thermal energy corresponding to a decrease in the apparent colloid mass, which inflates the SD profile. The effect of added salt is to reduce the entropic lift or, equivalently, to weaken the electric field: when counterions diffuse upwards in a salt-free dispersion, only colloids are available as anionic chaperons to maintain electroneutrality. However, at higher ionic strength, molecular anions also accompany counterions, whereas at excess salt, charge separation is suppressed and the electric field vanishes.

To quantify this qualitative explanation and analyse the non-barometric SD profiles (Fig. 1) we extend equation (1) with an electrical term. The resulting non-barometric formula also follows

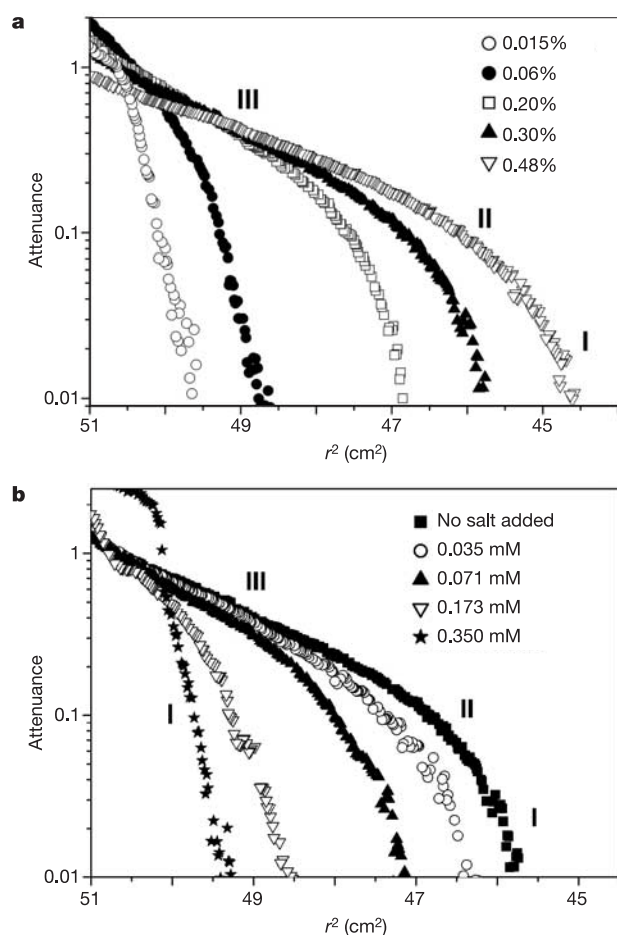


Figure 1 Experimental sedimentation–diffusion equilibrium profiles for dispersions of charged silica spheres in ethanol plotted as attenuation against the square of the radial distance from the centre of rotation. **a**, Representative profiles for dispersions having different initial volume fractions, with no added salt. Increase in the (very low) initial silica volume fraction significantly expands the exponential decay III as a consequence of the colloidal charge. **b**, Representative profiles for dispersions having a fixed initial silica volume fraction of 0.30%, at different salt concentrations. Increasing added amounts of LiNO₃ gradually compress the profile towards the barometric distribution I.

from the classical Donnan membrane equilibrium for an inhomogeneous colloidal solution^{5,13}. Here we give an equivalent but more concise derivation. Consider a mixture of three species: colloids with charge z and number density ρ , monovalent cations with concentration c_+ , and monovalent anions with concentration c_- . The dispersion is in thermodynamic equilibrium with an external salt reservoir with a total ion concentration of $2c_s$. The essential assumptions are that each separate species behaves ideally and that the three species jointly satisfy charge neutrality. For non-

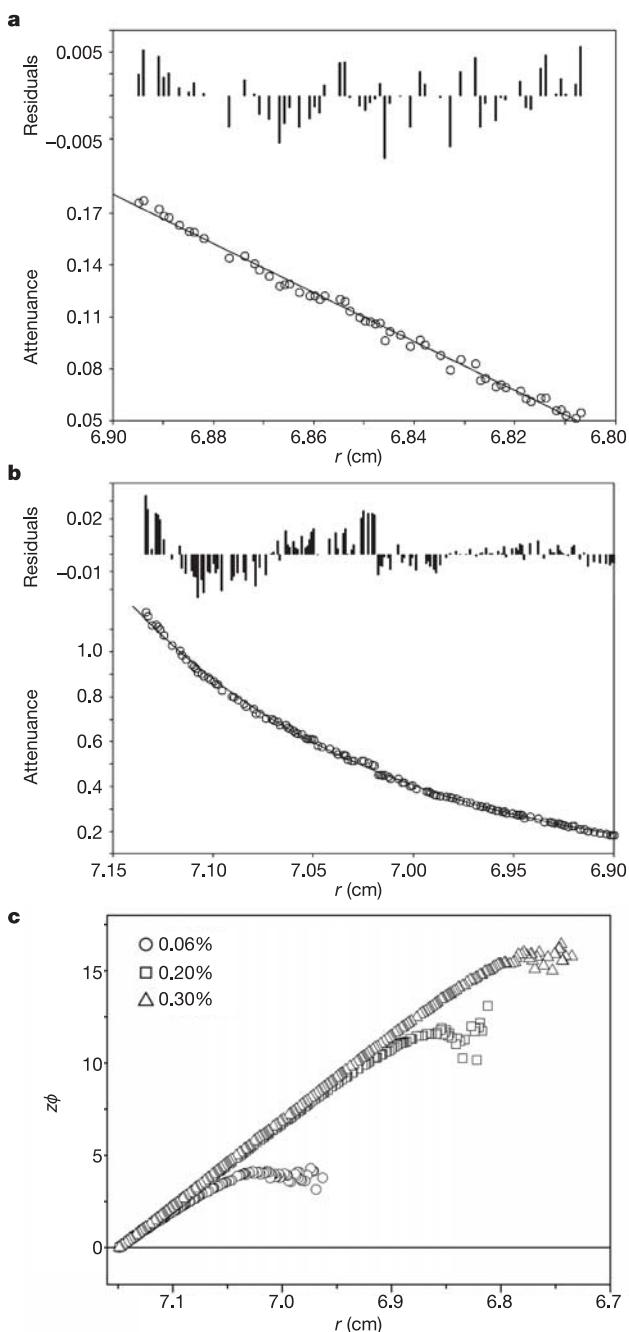


Figure 2 Analysis of the experimental profiles from Fig. 1a. **a**, An example of a nearly perfect quadratic fit to the data extracted from region II of the 0.30% sample, which clearly confirms the peculiar decay predicted for region II (see the text and Methods). **b**, An example of a nearly perfect exponential fit to the data extracted from region III of the same sample, resulting in a different centrifugal length from that of region I. **c**, Non-dimensional electrical potentials $z\phi$ plotted against the radial distance for three initial volume fractions of particles.

interacting, weightless ions, thermodynamic equilibrium requires the gradient in their osmotic pressure to balance the force (per volume) exerted on the ions by the internal electric field E :

$$-k_B T (dc_{\pm}/dx) \pm ec_{\pm} E = 0 \quad (2)$$

Here e is the proton charge; the electrostatic potential ψ follows from $E = -d\psi/dx$. The boundary condition on equation (2) is that $c_+ = c_- = c_s$ in the reservoir where $\psi = 0$. For the much heavier colloids the equilibrium condition contains the gravitational force per volume:

$$-k_B T (d\rho/dx) - z\rho E - m\rho g = 0 \quad (3)$$

again assuming non-interacting particles. The electric field term always accompanies the gravitational field term and vanishes only if gravity is switched off. To find the non-barometric SD profile from equation (3) one more equation is needed, for which we use here the macroscopic neutrality condition

$$-z\rho + c_+ - c_- = 0 \quad (4)$$

Integrating equation (2) yields the ion concentration profiles c_+ and c_- . Substituting them in equation (4) and combining the result with equation (3) finally yields the SD profile

$$\ln(y) + z \operatorname{arcsinh}(y) = -(x/L) + C \quad (5)$$

Here $y = z\rho/(2c_s)$ is a non-dimensional number density of colloids and C is an integration constant determined by the total number of colloids¹³. The reduced electrical potential $\phi = e\psi/(k_B T)$ is, according to equation (4), given by $\phi = -\operatorname{arcsinh}(y)$. The difference between the profiles in equations (1) and (5) is not merely adding the electrical work term; there is always an electrical field E opposing gravity whenever $z > 0$ (for explicit expressions of E see ref. 13). The homogeneous electric field E in regions II and III (see below) is expected to have a strength of order $E \approx mg/(ez)$. These two regions can be regarded as a macroscopic condenser^{6,13} resulting from the charge separation that produces the equilibrium macroscopic field. Note that equations (1) and (5) are one-dimensional SD profiles in the gravitational field; for profiles in the radial field of an ultracentrifuge see Methods. For $zy \ll 1$ (region I), equation (5) asymptotes to the barometric exponential (equation (1)), whereas for $y \gg 1$ (region III) the profile also decays exponentially, as $\ln(y) \sim -x/((z+1)L)$, but with an inflated gravitational length of $(z+1)L$ corresponding to a strongly reduced effective mass. Our experiments (Fig. 1) clearly show both exponential regions in one and the same SD profile. Moreover, when salt is added to decrease the average value of y , the barometric region I grows at the expense of region III, which eventually disappears (Fig. 1b). This marked salt effect confirms equation (5) and previous calculations⁶. Note that in Fig. 1b only little salt is needed to change SD profiles substantially. Without added salt, incidentally, the ion concentration, estimated from electrical conductivity measurements, is only of the order of 10^{-5} M, corresponding to a Debye screening length of about 30 nm.

In addition to the two exponential regions, a linear concentration decay is predicted⁶ in an intermediate region II, characterized by a low colloid density ($y \ll 1$) and a high counterion density ($zy \gg 1$), inequalities for which equation (5) indeed yields the linear asymptote: $\operatorname{arcsinh}(y) \sim y = 1 - x/(zL)$ (in the centrifugal field this decay is quadratic; see Methods). This most unusual SD

profile, ascribed to a near cancellation of gravity by the electric field⁶, is quite narrow unless z is very large. The silica sphere charge is only $z \approx 50$ (see Methods), but owing to the high resolution in our SD profiles a zoom-in (Fig. 2a) nevertheless clearly reveals the region II.

Equation (5) implies a significant rescaling of SD profiles, by extracting the barometric part (region I) from the experimental SD profiles to obtain $z\phi$ as a function of altitude. The results (Fig. 2c) show unambiguously the presence of a macroscopic electric field in a substantial part of SD profiles. The electrical potential decreases with the radial distance r and flattens on entering region I towards the constant electrical potential of the background electrolyte depleted of colloids. The largest overall potential jump in Fig. 2c is about 8 mV for $z \approx 50$, which agrees in order of magnitude with electrochemical Donnan potential measurements on silica dispersions in ethanol, with comparable differences in concentrations as in the SD profiles (B. Ern , unpublished observations). The slope $z d\phi/dr$ in Fig. 2c is proportional to the electrical force on colloids, which indeed almost compensates for the buoyant mass: for our silica spheres, the experimental slope (46.6 cm^{-1}) corresponds to a ratio of the electric force zeE to the centrifugal force $m\omega^2 r$ very close to unity. Note that for this calculation we do not need the value of z .

The applicability of equation (5) is noteworthy because it disregards all colloid–colloid and ion–ion correlations as well as electrical potential variations on the scale of colloids and ions. One would expect deficiencies in this model to show up in the more concentrated part of the profile (region III). We found for some of the profiles in Fig. 1a that the exponential decay in region III extends over a wider spatial range than predicted by equation (5). Moreover, a fit to region III (Fig. 2b) using the appropriate radial SD profile (see Methods) yields $z \approx 4$ –8 (Table 2). These values of z are low in comparison with values estimated from region II (Table 2) and electrophoresis (Table 1), which agree very well at lower volume fractions. Clearly, an extension of the theory beyond equation (5) is needed to further clarify the SD profiles at low altitude. This extension does not necessarily involve colloidal interactions (which might contribute to non-barometric behaviour): in our SD profiles these interactions have a much smaller effect than the electric field. We verified this by numerical calculations of the osmotic second virial coefficient for colloids interacting by means of an electrical double-layer repulsion^{5,10}: for concentrations in region III, the first-order correction in the virial expansion of the osmotic pressure is much smaller than the total pressure exerted by non-interacting colloids and their counterions. This confirms that also in region III the electric field is the dominant cause of non-barometric behaviour.

Thus, we have experimentally demonstrated the existence of a macroscopic electric field, producing strongly non-barometric SD profiles of charged colloids. The unusual shape of experimental SD profiles and the marked effect of added salt validate equation (5), which is based only on the assumptions of ideal species and charge neutrality. This simple model, which explains our results surprisingly well, nevertheless needs extension to quantify region III better. Although our experiments relate to charged silica spheres, it is clear that non-barometric behaviour can occur for any type of sedimenting charged species. In this respect it should be noted that the mass determination of biomolecules by centrifugal analysis at excess salt, to suppress charge effects, certainly makes sense in the light of our findings. However, our work shows clearly that an excess of salt also suppresses important information on the electrical properties of macromolecules: centrifugal analysis of polyelectrolytes at low salt concentration still provides the molecular mass (region I), but in principle also yields their valency and the macroscopic electrostatic potential that they generate in solution. It would be interesting to (re)examine the sedimentation of polyelectrolytes such as DNA or proteins from this perspective. □

Table 2 Results of fit analysis of profiles presented in Fig. 1a

Volume fraction (%)	$L_{\omega,1}$ (cm)	z_{II}	z_{III}
0.06	0.37	45	4
0.20	0.39	55	4
0.30	0.39	65	5
0.48	0.37	80	8

$L_{\omega,1}$ is the centrifugal length, determined from region I; z_{II} is the number of elementary charges on particles, determined from region II; z_{III} is the number of elementary charges on particles, determined from region III.

Methods

Silica dispersions

Amorphous silica spheres (Table 1) were synthesized by polymerization of (hydrolysed) tetraethoxysilane in an ethanol–ammonia mixture^{14,15}. The silica surface was modified by covalent reaction with 3-methacryloxypropyltrimethoxysilane (TPM) during distillation, simultaneously transferring the silica spheres to absolute (analytical grade) ethanol. The stock dispersion for the ultracentrifugation experiments was free from contaminants such as polymeric species or secondary silica particles, which might have affected SD profiles⁵. TPM-coated silica spheres do not aggregate in ethanol and are well documented model particles for the study of charge-stabilized colloids^{15,16}. Particle characterization (Table 1) was performed on the same samples as used for ultracentrifugation.

Ultracentrifugation

Ultracentrifugation was performed with a Beckman Optima XL-A analytical ultracentrifuge, in accordance with the manufacturer's instruction manual, in sector-shaped sedimentation cells thermostatically controlled at 298 K. The optical system measures the transmittance, H , of the sample against a reference of pure ethanol or salt solution and the results are converted to the attenuation $A = \ln(1/H)$ as plotted in Fig. 1. The attenuation is linearly proportional to the number density of silica particles up to a silica volume fraction of 2.5%, as checked with light attenuation (extinction) measurements. A silica sphere radius of about 22 nm was convenient for achieving profiles of sufficient thickness, which nevertheless decay fully within the available (radial) space to the solvent depleted of particles (within the limits of the sensitivity of the instrument). SD equilibrium was achieved on a time scale of up to 7 days at low rotor speeds (1100 r.p.m., corresponding to an acceleration of 88 g in the centre of the cell). Homogenization and recentrifugation of the same cell content reproduced the SD profiles very well, which also confirms that the silica spheres do not aggregate in the centrifugal field. The radial centrifugal field requires a modification of equation (5):

$$\ln(y) + z \operatorname{arcsinh}(y) = (r^2 - r_b^2)/(2L_\omega^2) + C \quad (6)$$

Here r is the radial distance from the rotor axis, r_b denotes the cell bottom and L_ω is the centrifugal length. The three regions are now characterized by the following equations: $\ln(y) \sim C_1 + (r^2 - r_b^2)/(2L_\omega^2)$ (region I), $y \sim C_2 + c_s(r^2 - r_b^2)/(z^2 L_\omega^2)$ (region II) and $\ln y \sim C_3 + (r^2 - r_b^2)/[2(z + 1)L_\omega^2]$ (region III), where r_1 is the crossover between the first two regions. SD profiles were always fitted with these quadratic arguments, although the experimental profiles are narrow so that they are very close to those formed in gravitational field (see, for example, Fig. 2a). For the radial electric field corresponding to equation (6) see ref. 13. The initial volume fractions of the homogeneous samples subjected to ultracentrifugation varied from 0.01% to 0.5%; after the SD equilibrium was reached, the altitude-dependent volume fraction of particles shows, at the bottom of profiles, values 7-fold down to 4-fold the initial values, respectively.

Electrophoretic mobility and electrical conductivity

Electrophoretic mobility and electrical conductivity were measured with a DELSA 440SX (Coulter) at 298 K for all samples. The mobility was determined from a Lorentzian fit to the measured peaks at four scattering angles and was further used to estimate the charge z of the particles in the frame of the Debye–Hückel model with a previously reported method¹⁷. Conductivities obtained from the DELSA were used to estimate the ion particle number density by the method of ref. 17. We found a negligible dependence of the mobility on sample volume fraction; that is, the measured values for all samples were $-0.95 \pm 0.5 \mu\text{m cm V}^{-1} \text{s}^{-1}$ (mean $\pm$ s.d.). The conductivity increased weakly with the volume fraction of particles, being in the range $1.4\text{--}2.9 \mu\text{S cm}^{-1}$.

Received 30 March; accepted 4 May 2004; doi:10.1038/nature02672.

- Harding, S. E., Rowe, A. J. & Horton, J. C. (eds) *Analytical Ultracentrifugation in Biochemistry and Polymer Science* (Royal Society of Chemistry, Cambridge, 1993).
- Stryer, L. *Biochemistry*, 4th edn (Freeman, New York, 1995).
- Svedberg, T. & Pederson, O. *Die Ultrazentrifuge* (Dresden, 1940).
- Piazza, R., Bellini, T. & Degiorgio, V. Equilibrium sedimentation profiles of screened charged colloids: a test of the hard-sphere equation of state. *Phys. Rev. Lett.* **71**, 4267–4270 (1993).
- Philipse, A. P. & Koenderink, G. H. Sedimentation-diffusion profiles and layered sedimentation of charged colloids at low ionic strength. *Adv. Colloid Interface Sci.* **100–102**, 613–639 (2003).
- Van Roij, R. Defying gravity with entropy and electrostatics: sedimentation of charged colloids. *J. Phys.: Condens. Matter* **15**, S3569–S3580 (2003).
- Hynninen, A. P., Van Roij, R. & Dijkstra, M. Sedimentation profiles of charged colloids: Entropic lift and charge separation. *Europhys. Lett.* **65**, 719–725 (2004).
- Biben, T. & Hansen, J. P. Sedimentation equilibrium in concentrated charged-stabilized colloidal suspensions. *J. Phys.: Condens. Matter* **6**, A345–A349 (1994).
- Löwen, H. The apparent mass in sedimentation profiles of charged suspensions. *J. Phys.: Condens. Matter* **10**, L479–L485 (1998).
- Hunter, R. J. *Foundations of Colloid Science* Vol I (Clarendon, Oxford, 1991).
- Becker, W. M., Kleinsmith, L. J. & Mardin, J. *The World of the Cell* (Benjamin Cummings, San Francisco, 2000).
- Meselson, M. & Stahl, F. W. The replication of DNA in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **44**, 671–682 (1958).
- Philipse, A. P. Remarks on the Donnan condenser in the sedimentation–diffusion equilibrium of charged colloids. *J. Phys.: Condens. Matter* (submitted).
- Philipse, A. P. & Vrij, A. Preparation and properties of nonaqueous model dispersions of chemically modified charged silica spheres. *J. Colloid Interf. Sci.* **128**, 121–153 (1989).
- Philipse, A. P. & Vrij, A. Determination of static and hydrodynamic interactions between monodisperse silica spheres in an optically matching, organic solvent. *J. Chem. Phys.* **88**, 6459–6470 (1988).

- Thies-Weesie, D. M. E., Philipse, A. P., Nägele, G., Mandl, B. & Klein, R. Non-analytical concentration dependence of sedimentation of charged silica spheres in an organic solvent. Experiments and calculations. *J. Colloid Interf. Sci.* **176**, 43–54 (1995).
- Rasa, M., Philipse, A. P. & Meeldijk, J. D. Heteroaggregation, reprecipitation, and stability in mixtures of oppositely charged colloids. *J. Colloid Interf. Sci.* (in press).

Acknowledgements We thank R. van Roij and G. Koenderink for discussions and collaboration, B. Ern  for sharing insights and unpublished results on Donnan potentials, N. Zuiverloon for synthesizing the silica spheres, and D. Thies-Weesie for maintenance of the ultracentrifuge. This work was supported financially by The Netherlands Organization for Scientific Research (NWO/Stichting Chemische Wetenschappen).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.P.P. (a.p.philipse@chem.uu.nl).

Role of CO₂ in the formation of gold deposits

G. N. Phillips¹ & K. A. Evans²

¹CSIRO Division of Exploration and Mining, PO Box 3, Central Park, Victoria 3145, Australia

²CSIRO Division of Exploration and Mining, School of Geosciences, Building 28, Monash University, Victoria 3800, Australia

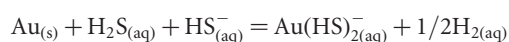
Much of global gold production has come from deposits with uneconomic concentrations of base metals, such as copper, lead and zinc¹. These ‘gold-only’ deposits are thought to have formed from hot, aqueous fluids rich in carbon dioxide², but only minor significance has been attached to the role of the CO₂ in the process of gold transport. This is because chemical bonding between gold ions and CO₂ species is not strong³, and so it is unlikely that CO₂ has a direct role in gold transport. An alternative indirect role for CO₂ as a weak acid that buffers pH has also appeared unlikely, because previously inferred pH values for such gold-bearing fluids are variable^{2,4,5,6}. Here we show that such calculated pH values are unlikely to record conditions of gold transport, and propose that CO₂ may play a critical role during gold transport by buffering the fluid in a pH range where elevated gold concentration can be maintained by complexation with reduced sulphur. Our conclusions, which are supported by geochemical modelling, may provide a platform for new gold exploration methods.

The world's major gold-only deposits have formed by precipitation of gold from aqueous fluids migrating on a kilometre-scale through the bedrock that comprises the upper part of the Earth's crust. These fluids show a remarkably consistent set of characteristics. They are at temperatures above 200 °C, rich in CO₂ (up to 20–30 mol%; Table 1), sulphur-bearing, low salinity compared to most ore-bearing fluids, and have a redox state more reducing than the haematite–magnetite buffer such that the sulphur is predominantly in the reduced state¹. These characteristics have been confirmed (by techniques including fluid-inclusion and stable-isotope analyses) for deposit types that account for the bulk of the world's gold production, both today and historically—for example, Archaean greenstone gold^{2,7}, slate-belt gold⁸, the Mother Lode of California⁶, gold in banded iron formations², Carlin-type gold⁹ and the Witwatersrand goldfields^{9,10}.

With the exception of the high CO₂ concentrations, these shared fluid characteristics can be explained by the chemical attributes of the dissolved gold. The somewhat reducing fluids and elevated temperatures favour the ionic state of Au⁺ for the gold in solution³.

Au^+ is the softest of cations, that is, it prefers covalent bonding^{3,11,12}. Soft cations preferentially bond with soft anions and ligands, so Au^+ forms strong complexes with soft ligands. HS^- is one of the few ligand candidates to meet the dual criteria of being moderately abundant in these fluids, and soft in character. Hydrosulphide complexing of gold is consistent with laboratory experiments¹³, with inferred sulphidation and redox depositional mechanisms in the formation of gold deposits^{11,14}, and with the abundance of sulphide minerals (such as pyrite) in gold deposits. Hydrosulphide complexing of gold is also consistent with low salinities. Chlorine forms a moderately hard anion (Cl^-), and although it can complex with gold under oxidizing conditions, under the more reducing conditions inferred for gold-only deposits, it bonds preferentially with harder cations, such as Cu, Pb and Zn. Low salinities and a lack of these base metals in gold-only deposits are thus consistent with hydrosulphide complexing.

To understand the role of the CO_2 , it is necessary to examine interactions between sulphur complexation and other fluid constituents during gold transport and deposition. The concentration of gold in potentially ore-forming solutions is a function of the distribution of sulphur species. For temperatures of 200–400 °C, pressures of 200 MPa and near-neutral pH, $\text{Au}(\text{HS})_2^-$ is the dominant gold–hydrosulphide complex¹⁵. Gold concentrations can be related to those of the sulphide species via:



K , the stability constant, is defined by:

$$K = \frac{\{\text{Au}(\text{HS})_2^-\} \{\text{H}_2\}^{0.5}}{\{\text{H}_2\text{S}\} \{\text{HS}^-\}}$$

where {...} indicates activity. This can be rearranged to give $\{\text{Au}(\text{HS})_2^-\}$, the gold activity in solution:

$$\{\text{Au}(\text{HS})_2^-\} = \frac{K \{\text{H}_2\} \{\text{HS}^-\}}{\{\text{H}_2\}^{0.5}}$$

The maximum gold solubility at a given hydrogen activity occurs when the product $\{\text{H}_2\} \{\text{HS}^-\}$ is maximized, and as these sulphur species are related by $\text{H}_2\text{S} \leftrightarrow \text{HS}^- + \text{H}^+$, the maximum occurs when $\{\text{H}_2\text{S}\} = \{\text{HS}^-\}$. For $K_1(\text{H}_2\text{S}) = \{\text{HS}^-\} \{\text{H}^+\} / \{\text{H}_2\text{S}\} = \{\text{H}^+\}$, it follows that the maximum occurs where $\text{pH} = \text{p}K_1(\text{H}_2\text{S})$.

With respect to pH and redox parameters, the maximum gold solubility occurs close to the point where hydrogen sulphide, bisulphide and sulphate species coexist in equal quantities. The redox state must be largely wallrock-controlled¹⁴, because the fluid has a low redox buffering capacity. However, further information is required to determine whether pH is controlled by fluid or wallrock. Two lines of evidence suggest that pH during gold transport is fluid-buffered. First, in the strongly altered wallrocks typical of goldfield environments, mineral buffers of pH are often exhausted^{14,16}, suggesting that pH was controlled by components within the fluid. Second, pH changes are not usually implicated as the main control on gold deposition, indicating that pH in the fluid was largely independent of wallrock.

We propose that CO_2 within the fluid provided the buffering capacity needed to maintain elevated gold solubilities. A pH buffer needs to be both abundant and to have weak acid–base properties. H_2CO_3 is a weak acid that is abundant in the gold-bearing fluid (up to 200,000–300,000 p.p.m. total for all carbon species), and has pH buffering capability through $\text{H}_2\text{CO}_3 = \text{HCO}_3^- + \text{H}^+$ relationships. Importantly, the first dissociation constant for H_2CO_3 is close to the first dissociation constant of H_2S for a range of geologically relevant temperatures (Fig. 1a). We have already shown that maximum gold solubility occurs where the pH equals $\text{p}K_1(\text{H}_2\text{S})$ for any given hydrogen activity, so that a fluid experiencing a significant degree of CO_2 buffering near $\text{p}K_1(\text{H}_2\text{CO}_3)$ would automatically be maintained in a region of elevated gold solubility.

Two issues militate against acceptance of this role for CO_2 . The first is that published estimates of pH in auriferous fluids are highly variable, and often significantly different to $\text{p}K_1$ for CO_2 or H_2S (Table 1). This variability is incongruous with the remarkably consistent chemical characteristics of gold-only deposits¹. These pH estimates rely on K^+/H^+ and Na^+/H^+ ratios derived from

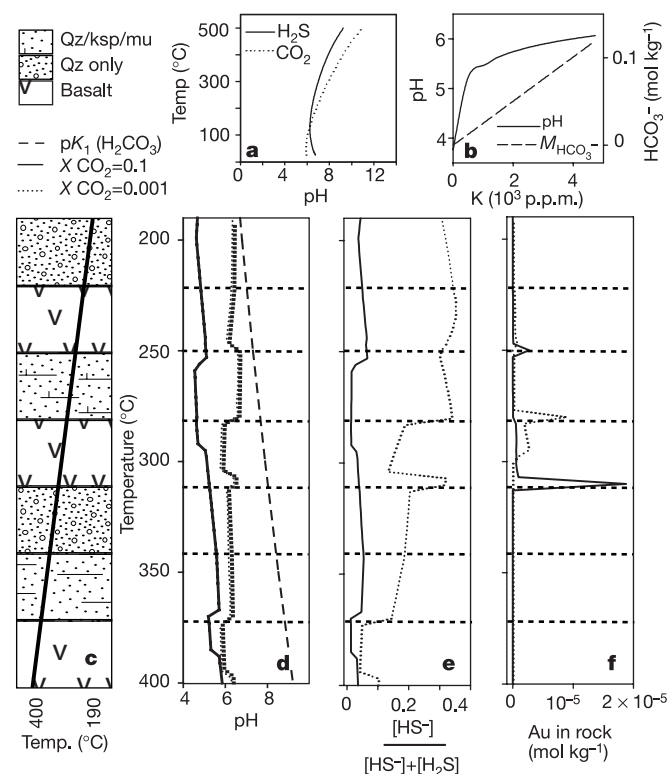


Figure 1 Results of HCh modelling in the K-Fe-Mg-Ca-Al-Si-C-H-Au-S-O system. All figures are constructed using HCh and the Unitherm database. Pressure is 200 MPa in all cases. Qz, quartz; mu, muscovite; ksp, K-feldspar. **a**, Temperature–pH plot showing $\text{p}K_1$ for weak acids H_2S and H_2CO_3 . $\text{p}K_1$ for H_2CO_3 is very close to $\text{p}K_1$ of H_2S for a wide range of geological temperatures. Maximum gold solubility occurs along the H_2S line. **b**, pH– K plot for a H_2O – CO_2 fluid (15 mol% CO_2) at 350 °C and 200 MPa. $M_{\text{HCO}_3^-}$ is the molality of the HCO_3^- ion. Without K , a fluid in this system is quite acid, but for K -present conditions, as implicated in the formation of gold deposits, the pH is much closer to $\text{p}K_1(\text{H}_2\text{S})$. The concentration of HCO_3^- , and hence the buffering capacity of the fluid, have a positive correlation with the K concentration. **c**, Sequence of rock types used for the flow-through reactor modelling shown in **d–f**. Fluids equilibrate with a gold and sulphur-bearing basalt (not shown), and then cool from 400 °C to 190 °C as they pass upwards through successive rock types. Time-integrated fluid:rock ratios by weight are 50. **d–f**, Results of flow-through reactor modelling. Solid line, 10 mol% CO_2 ; dotted line, 0.1 mol% CO_2 . **d**, pH of the modelled fluids as they progress through the rock sequence. The pH of the CO_2 -bearing fluid (initially 10 mol% CO_2) decreases with temperature and is virtually parallel to the dashed line which shows $\text{p}K_1(\text{H}_2\text{CO}_3)$. The pH of the essentially CO_2 -free fluid (initially 0.1 mol% CO_2) diverges markedly from the trend of $\text{p}K_1(\text{H}_2\text{CO}_3)$. Steps in pH indicate the positions of reaction fronts in the wallrocks. **e**, Proportion of the sulphur as HS^- (defined as $[\text{HS}^-]/([\text{HS}^-] + [\text{H}_2\text{S}])$). In the CO_2 -bearing fluid there are small changes related to changes of pH in **d**. The proportion of HS^- is relatively constant because of the CO_2 -buffering of pH. This is strong confirmation that pH buffering by CO_2 controls the distribution of sulphur species. For the low CO_2 fluid, the proportion of HS^- increases with decreasing temperature reflecting the divergence from $\text{p}K_1(\text{H}_2\text{CO}_3)$ in **d**. Total reduced sulphur concentrations in both cases range from 20 mM at the base of the column to 0.01 mM at the top. **f**, Gold deposited from the fluids modelled in **c** and **d**. The CO_2 -bearing fluid deposits the gold in a more focused way in the Fe-rich basalt in response to sulphidation reactions. The CO_2 -poor fluid deposits gold over a broader part of the column.

observed alteration mineral assemblages (for example, pyrophyllite plus muscovite), and, in some cases, measurement of the K/Na ratio in fluid inclusions. Such calculations of pH values are likely to be unreliable because extensive and widespread K and Na metasomatism observed in gold deposit wallrocks indicates that both elements are intimately involved in the depositional process. It is therefore likely that these pH values derived indirectly using measurements of Na and K relate to intermediate depositional processes rather than values during gold transport.

The second issue is based on the fact that in a pure C-O-H-S fluid, the only cation species would be H^+ , and pH would be significantly below the pK_1 for either CO_2 or H_2S . Under these circumstances, both H_2S and CO_2 would be almost fully associated, and could not buffer pH. However, ore-forming fluids interact with wallrocks as they travel through the Earth's crust, and H^+ ions exchange for Na^+ and K^+ ions. Loss of H^+ through this exchange increases the pH of the fluid to the vicinity of $pK_1(H_2CO_3)$ for geologically reasonable K^+ concentrations (Fig. 1b), so this issue is not an obstacle to the suggestion that CO_2 buffers pH.

We have tested the hypothesis that CO_2 buffering can control pH in gold-bearing fluids using the geochemical modelling program HCh and its associated Unitherm database. These comprise a software package designed for equilibrium modelling of fluid-rock systems¹⁷. The chemical system modelled was K-Mg-Fe-Al-Ca-C-Si-H-O-Au-S, and the database includes Au^+ and Au^{3+} hydroxides and hydrosulphides (see Methods section for full list). Chloride was not included because there is no evidence for geologically high salinity fluids in the deposits of interest (Table 1). A brief conceptual description of the model is given here, and further details can be found in the Methods section. H_2O - CO_2 fluids with 10 mol% CO_2 were initially equilibrated with a basalt at 200 MPa, 400 °C, and a fluid:rock ratio of 1:1. Equilibrated fluids were then added to the bottom of a modelled rock column (Fig. 1c). Each of the seven lithological units within the column is composed of 10 reacting cells; the fluid equilibrates with each cell before being passed on to the next. Fluid-rock ratios were set to 5:1 to mimic the focusing of fluids in ore-forming systems. Multiple waves of fluid passed up the hypothetical column; initial fluids for each wave were equilibrated with the gold-bearing basalt, as for the first step, but were then reacted with cells of rock that had already been modified by fluids from the preceding steps. Temperature was set to decrease by 3 °C per cell as the column was ascended. The redox state was set above the pyrrhotite-pyrite-magnetite buffer, such that neither methane nor sulphate were stable with respect to carbon dioxide and hydrogen sulphide, respectively. Results were compared with a control run with 0.1 mol% CO_2 . The model is used to derive trends and relative, rather than absolute, values, because of inherent uncertainties in thermodynamic data and the limited applicability of the thermodynamic model to CO_2 -rich solutions.

Selected results of the final wave of the model are shown in Fig. 1d-f, which illustrate pH, proportion of the sulphur as HS^- for CO_2 -bearing and CO_2 -free fluids, and gold concentration in the

rock, respectively. pH in the CO_2 -bearing fluid is lower than that of CO_2 -free (Fig. 1d), because of the acidic nature of the CO_2 . The CO_2 -free fluid can carry more gold, but this is not a limiting factor at the source, as typical basaltic source rocks contain only a few p.p.b. of gold, and devolatilization results in gold undersaturated solutions over a wide range of temperature. Changes in sulphur species distribution (Fig. 1e) are almost eliminated in the CO_2 -bearing fluid, providing a stable chemical environment for gold transport through the Earth's crust. Modelled mineral abundances (not shown) indicate that CO_2 -bearing and CO_2 -free fluids will deposit gold in basalt as a result of pyrite-forming reactions between sulphur in the fluid, and iron within the rock. This mechanism is important in gold deposits¹⁴, and its occurrence here supports the CO_2 buffering hypothesis.

A significant difference between the two hypothetical fluids lies in the distribution of deposited Au (Fig. 1f). The CO_2 -bearing fluid focuses gold deposition over a narrow distance and temperature range, whereas the CO_2 -free fluid deposits gold over a greater distance but at lower concentrations. The former is more likely to result in an economic gold deposit. The contrasting deposition pattern results from CO_2 buffering pH to lower levels as temperature decreases (Fig. 1d), leading to a greater deposition of Au when suitable rock types (for example, Fe-rich basalts) are encountered. The difference between CO_2 -bearing and CO_2 -free fluids is enhanced by a basalt-induced pH increase in the CO_2 -free fluid. The increase enhances gold solubility, reducing deposition by sulphidation, whereas the CO_2 -bearing fluid is buffered and so does not experience this pH increase, and deposits a greater proportion of the dissolved gold.

Our arguments do not incorporate the potentially significant effects of CO_2 on hydrogen-bonding in water and consequent changes in ion hydration, complex stability, mineral solubility and aqueous speciation¹⁸. Neither do we consider the effect of H_2O - CO_2 immiscibility; this may be important in gold deposition, but is unlikely to play a significant role in transport.

To summarize, the role of CO_2 has commonly been regarded as peripheral because carbonate alteration and economic gold are only approximately coincident, and theory does not suggest gold complexing with the hard carbonate ligand. We suggest that CO_2 is essential in the formation of gold-only provinces and deposits, via its abilities to buffer fluid pH at a level that can maintain elevated concentrations of gold in solution, and allow gold deposition in chemically favourable host rocks. Previous estimates of the pH of gold-bearing fluids are likely to reflect transient depositional processes rather than the gold-transporting stage.

Our results have important implications for gold exploration. Process-based genetic models for gold deposits have already proved useful in the Yandal gold province of Western Australia, where 10 million ounces of Au have been discovered since 1990¹⁹. Better genetic models will help prioritize exploration parameters. Here we show that effective Au transport requires both H_2S and CO_2 , which must, implicitly, be incorporated at the source of the Au, constrain-

Table 1 Gold-related fluid characteristics from major ore deposits

	CO_2 (mol%)	pH evidence	Inferred pH	Salinity (wt% NaCl eq.)
Witwatersrand ⁴	NA	Pyrophyllite	Acid	NA
Witwatersrand ²⁵	NA	Muscovite	3.8–5	12
Vaal Reefs South, South Africa ¹⁰	Significant	NA	NA	6–18
Archaean Greenstone (general) ⁵	NA	Dolomite	Near-neutral	NA
Hollinger mine, Canada ⁷	10–20	NA	NA	Low to variable
Hunt, Kambalda ^{2,14}	25	Na/K	6.9	2
Slate Belt (general) ²⁵	NA	Calcite	Near-neutral	NA
Wattle Gully, Castlemaine, Victoria ⁸	20	NA	NA	Low
Carlin deposit, Nevada, USA ⁹	5–10	NA	NA	3
Mother Lode, California ⁶	10	Na/K	6	3

Fluid compositions refer to the majority of recorded fluids, rather than to the full range. Gold deposits with economic base metals may have salinities of 50 wt% NaCl equivalent or greater^{26,27}. NA, not available.

ing the genesis of ore-bearing fluids. Characteristics determined at the source will be consistent throughout a group of gold deposits, for example, Yandal gold province, whereas those related to depositional sites, such as high iron or carbon content, will account for some of the variability between deposits. □

Methods

HCh description

The HCh program package was developed for geochemical modelling of rock–fluid systems at moderate to high temperatures (1,000 °C) and moderate pressures (<500 MPa). It includes several programs that interact to calculate equilibrium compositions of multi-phase chemical systems using free-energy minimization techniques, and a thermodynamic database (Unitherm). This database, although carefully compiled, is taken from a variety of sources and is not internally consistent; however, observed compositional trends and comparisons are likely to be valid. Free energies of solid phases are calculated conventionally using reference entropies, enthalpies and volumes, with a Maier–Kelley expression for heat capacity¹⁷. Molar volumes of solids are assumed to be independent of pressure. The properties of pure water are calculated using the Haar–Gallager–Kell model²⁰. Properties of selected aqueous endmembers (primary species) are calculated via the revised Helgeson–Kirkham–Flowers equations of state^{21,22}. Other endmembers (secondary species) were calculated using values for the primary species and the modified Ryzenko–Bryzgalin model¹⁷. Apparent molal standard state Gibbs free energies at the pressure and temperature of interest, calculated by Unitherm, are given in Supplementary Table M1 for the principal aqueous species involved in reactions plus additional gold species. The standard state for aqueous species is unit activity in an ideal hypothetical 1 molal solution at the pressure and temperature of interest, referenced to infinite dilution. The standard state for solid phases and water is unit activity for the pure phase at the pressure and temperature of interest. Solid phases were assumed to be pure endmembers. Activity coefficients of aqueous charged species were calculated using the extended Debye–Hückel equation. Extended term parameters are available for a variety of background electrolytes^{23,24}. Neutral aqueous species were considered to mix ideally. The validity of the activity model is restricted to ionic strengths of less than 0.1; our modelled ionic strengths reach 0.14 at low temperatures, but are generally below 0.1. A greater concern results from the limited ability of the thermodynamic formulation to fully incorporate the effects of high CO₂ concentrations.

Modelled system

Modelling was performed within the system K–Fe–Mg–Al–Si–C–Ca–H–O–Au–S. Bulk compositions for the rock types described in the text are given in Supplementary Table M2. Rock compositions present are basalt, aluminous psammite and pure quartz. These rock compositions reflect those present in the Eastern Goldfields of Australia. High SiO₂ values in Supplementary Table M2 are artefacts of the normalization scheme used. Quartz is in excess in any case, and high values do not affect conclusions. Au contents were 100 p.p.b.; this is high, but did not lead to a Au-saturated fluid. Unitherm contains data for 71 potential mineral phases and 61 different aqueous species within this composition space. Aqueous gold-bearing species were Au⁺, AuOH, Au(OH)₂[−], AuHS, Au(HS)₂[−], Au³⁺ and Au₂(HS)₂S[−]. 15 of the 71 minerals were stable in at least one rock type during modelling. Mineral assemblages were consistent with those observed in gold-only deposits (Supplementary Table M3), given the limitations of bulk composition and constraints on mineral compositions. Various combinations of fluid:rock ratio and number of fluid waves were tested to assess the sensitivity of the model to the discretization scheme. Discretization affects the details, but not the overall trends and conclusions produced by the model. A fluid:rock ratio of 5:1 was chosen for the calculations shown in Fig. 1.

Received 8 January; accepted 10 May 2004; doi:10.1038/nature02644.

- Phillips, G. N. & Powell, R. Link between gold provinces. *Econ. Geol.* **88**, 1084–1098 (1993).
- Ho, S. E., Groves, D. I. & Phillips, G. N. Fluid inclusions as indicators of the nature and source of ore fluids and ore depositional conditions for Archaean gold deposits of the Yilgarn Block, Western Australia. *Trans. Geol. Soc. S. Afr.* **88**, 149–158 (1985).
- Puddephatt, R. J. *The Chemistry of Gold* (Elsevier, Amsterdam, 1978).
- Barnicoat, A. C. *et al.* Hydrothermal gold mineralization in the Witwatersrand basin. *Nature* **386**, 820–824 (1997).
- Morrison, G. W., Rose, W. J. & Jaireth, S. Geological and geochemical controls on the silver content (fineness) of gold in gold–silver deposits. *Ore Geol. Rev.* **6**, 333–364 (1991).
- Bohlke, J. K. Comparison of metasomatic reactions between a common CO₂-rich vein fluid and diverse wall rocks: intensive variables, mass transfers, and Au mineralization at Alleghany, California. *Econ. Geol.* **84**, 291–327 (1989).
- Smith, T. J., Cloke, P. L. & Kesler, S. E. Geochemistry of fluid inclusions from the McIntyre–Hollinger gold deposit, Timmins, Ontario. *Econ. Geol.* **79**, 1265–1285 (1984).
- Cox, S. F., Wall, V. J., Etheridge, M. A. & Potter, T. F. Deformational and metamorphic processes in the formation of mesothermal vein-hosted gold deposits — examples from the Lachlan Fold Belt in central Victoria, Australia. *Ore Geol. Rev.* **6**, 391–423 (1991).
- Kuehn, C. A. & Rose, A. W. Carlin gold deposits, Nevada: origin in a deep zone of mixing between normally pressured and overpressured fluids. *Econ. Geol.* **90**, 17–36 (1995).
- Phillips, G. N., Klemd, R. & Robertson, N. S. Summary of some fluid inclusion data from the Witwatersrand Basin and surrounding granitoids. *Mem. Geol. Soc. India* **11**, 59–65 (1988).
- Phillips, G. N. & Groves, D. I. The nature of Archaean gold-bearing fluids as deduced from gold deposits of Western Australia. *J. Geol. Soc. Austr.* **30**, 25–39 (1983).
- Ahrland, S., Chatt, J. & Davies, N. R. The relative affinities of ligand atoms for acceptor molecules and ions. *Q. Rev. Chem. Soc.* **12**, 265–276 (1958).

- Seward, T. M. Thio-complexes of gold in hydrothermal ore solutions. *Geochim. Cosmochim. Acta* **37**, 379–399 (1973).
- Neall, F. B. & Phillips, G. N. Fluid-wall interaction in an Archean hydrothermal gold deposit: a thermodynamic model for the Hunt Mine, Kambalda. *Econ. Geol.* **82**, 1679–1694 (1987).
- Benning, L. G. & Seward, T. M. Hydrosulphide complexing of Au(I) in hydrothermal solutions from 150–400 degrees C and 500–1500 bar. *Geochim. Cosmochim. Acta* **60**, 1849–1871 (1996).
- Phillips, G. N. & Law, J. D. M. Witwatersrand gold fields: geology, genesis and exploration. *SEG Rev.* **13**, 439–500 (2000).
- Shvarov, Y. & Bastrakov, E. *HCh: A Software Package for Geochemical Equilibrium Modelling. User's Guide Record 125* (Australian Geological Survey Organisation, Canberra, 1999).
- Gibert, E., Moine, B., Schott, J. & Dandurand, J.-L. Modelling of the transport and deposition of tungsten in the scheelite-bearing calc-silicate gneisses of the Montagne Noire, France. *Contrib. Mineral. Petrol.* **112**, 371–384 (1992).
- Phillips, G. N. & Vearncombe, J. R. Exploration of the Yandal gold province, Yilgarn Craton, Western Australia. *CSIRO Explores* **1**, 1–26 (2003).
- Kestin, J., Sengers, J. V., Kamgar-Parsi, B. & Levelt Sengers, J. M. H. Thermophysical properties of fluid H₂O. *J. Phys. Chem. Ref. Data* **13**, 175–183 (1984).
- Tanger, J. C. & Helgeson, H. C. Calculation of the thermodynamic and transport-properties of aqueous species at high-pressures and temperatures — revised equations of state for the standard partial molal properties of ions and electrolytes. *Am. J. Sci.* **288**, 19–98 (1988).
- Shock, E. L., Oelkers, E. H., Johnson, J. W., Sverjensky, D. A. & Helgeson, H. C. Calculation of the thermodynamic properties of aqueous species at high pressures and temperatures. *J. Chem. Soc. Faraday Trans.* **88**, 803–826 (1992).
- Helgeson, H. C., Kirkham, D. H. & Flowers, G. C. Theoretical prediction of the thermodynamic behaviour of aqueous electrolytes at high pressures and temperatures: IV Calculation of activity coefficients, osmotic coefficients, and apparent molal and standard and relative partial molal properties to 600 degrees C and 5 kbar. *Am. J. Sci.* **281**, 1259–1516 (1981).
- Oelkers, E. H. & Helgeson, H. C. Triple-ion anions and polynuclear complexing in supercritical electrolyte-solutions. *Geochim. Cosmochim. Acta* **54**, 727–738 (1990).
- Robb, L. J. & Meyer, F. M. A contribution to recent debate concerning epigenetic versus syngenetic mineralization processes in the Witwatersrand basin. *Econ. Geol.* **86**, 396–401 (1991).
- Eastoe, C. J. A fluid inclusion study of Panguna porphyry copper deposit, Bougainville, Papua New Guinea. *Econ. Geol.* **73**, 721–748 (1978).
- Goellnicht, N. M., Groves, D. I., McNaughton, N. J. & Dimo, G. An epigenetic origin for the Telfer gold deposit. *Econ. Geol. Monogr.* **6**, 151–167 (1989).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Law and M. Hughes for critical comments; R. Smith for earlier discussions relating to the chemical behaviour of gold; and R. Phillips and S. Wood for comments and suggestions that helped to improve the manuscript. K.A.E. acknowledges honorary positions at Monash and Melbourne Universities.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.N.P. (neil.phillips@csiro.au).

Global biodiversity patterns of marine phytoplankton and zooplankton

Xabier Irigoien¹, Jef Huisman² & Roger P. Harris³

¹AZTI, Arrantzeta eta Elikagintzarako Institutu Teknologikoa, Herrera Kaia portualdea, 20110 Pasaia, Spain

²Aquatic Microbiology, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Nieuwe Achtergracht 127, 1018 WS Amsterdam, The Netherlands

³Plymouth Marine Laboratory, Prospect Place, Plymouth PL1 3DH, UK

Although the oceans cover 70% of the Earth's surface, our knowledge of biodiversity patterns in marine phytoplankton and zooplankton is very limited compared to that of the biodiversity of plants and herbivores in the terrestrial world. Here, we present biodiversity data for marine plankton assemblages from different areas of the world ocean. Similar to terrestrial vegetation^{1–3}, marine phytoplankton diversity is a unimodal function of phytoplankton biomass, with maximum diversity at intermediate levels of phytoplankton biomass and minimum diversity during massive blooms. Contrary to expectation, we did not find a relation between phytoplankton diversity and zooplankton

diversity. Zooplankton diversity is a unimodal function of zooplankton biomass. Most strikingly, these marine biodiversity patterns show a worldwide consistency, despite obvious differences in environmental conditions of the various oceanographic regions. These findings may serve as a new benchmark in the search for global biodiversity patterns of plants and herbivores.

The biodiversity of plants and herbivores often depends on ecosystem productivity. The available evidence suggests that several diversity–productivity patterns are possible^{4,5}, that these patterns change with the scale of observation^{4,6}, and that these patterns depend on the history of the community assembly⁷. Many studies have revealed a unimodal pattern, with maximal diversity at intermediate levels of productivity^{1–5,8–12}. Other studies revealed a monotonic increase of diversity with productivity^{4,5}. The majority of studies have focused on terrestrial plant communities and freshwater ecosystems, however. Although the oceans cover much more of the Earth's surface than does land, our knowledge of marine biodiversity patterns is still very limited¹².

We have compiled data on marine plankton assemblages, in terms of species composition and biomass, from 12 globally distributed areas. These areas are the Norwegian Sea, the North Atlantic Ocean, the Iceland Basin, the Irminger Sea, Long Island Sound, the North Sea, the English Channel, the Benguela and Oregon upwellings, the Indian Ocean, mesocosms in the Bergen fjord, and five meridional transects from 48°N to 50°S in the Atlantic Ocean. For all areas, data were collected on the species composition of nanophytoplankton (2–20 µm), microphytoplankton (20–200 µm), and microzooplankton (20–200 µm; heterotrophic dinoflagellates and ciliates). We call this our 'global data set'. For one of the Atlantic meridional transects, AMT4, we had data on picophytoplankton (0.2–2 µm; including *Prochlorococcus*, *Synechococcus* and picoeukaryotes) as well. For the English Channel we had data on mesozooplankton (200–20,000 µm; mainly copepods). Therefore, the AMT4 data and English Channel data were analysed separately from the global data set.

Recent studies indicate that microzooplankton are the main consumers of primary production in the oceans¹³. Microzooplankton biomass is an increasing saturating function of phytoplankton biomass (Fig. 1a). In areas with low phytoplankton biomass, microzooplankton biomass is on average ~20% of phytoplankton biomass. With increasing phytoplankton biomass, microzooplankton biomass saturates, suggesting that grazing intensity is on average less intense in areas with dense phytoplankton populations. This pattern is not restricted to microzooplankton. The same pattern is observed when mesozooplankton are included, as in the English Channel data (Fig. 1b).

The amazing biodiversity of plants in tropical rainforests is accompanied by a similarly astonishing biodiversity of animals. However, contrary to expectation, we observed only a very weak relation between phytoplankton diversity and microzooplankton diversity (Fig. 1c). Zooplankton diversity is obviously higher when mesozooplankton are included as well. With inclusion of mesozooplankton, however, we observed the same lack of relation between phytoplankton diversity and zooplankton diversity (Fig. 1d). That is, marine pelagic environments with a high plant diversity do not necessarily have a high herbivore diversity. Possible explanations for this unexpected difference between aquatic and terrestrial biodiversity patterns might be that many terrestrial herbivores (for example, many insects) are specialized on their host plants whereas zooplankton species are generally less specialized, or that the structural diversity of terrestrial vegetation offers many more spatial niches for herbivores than the unstructured aquatic habitat.

Phytoplankton diversity is a unimodal function of phytoplankton biomass (Fig. 2). Average phytoplankton cell biomass increased more than one order of magnitude with increasing phytoplankton

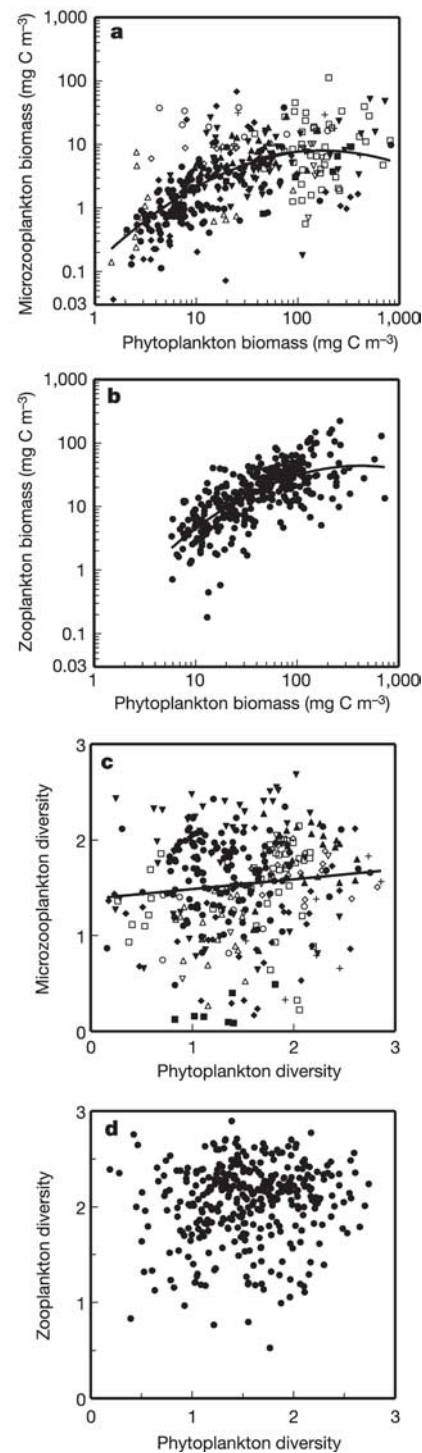


Figure 1 General characteristics of the data sets. **a**, Microzooplankton biomass versus phytoplankton biomass. Global data set, multiple regression: $\log[y] = -0.89 + 1.55\log[x] - 0.33(\log[x])^2$; $R^2 = 0.44$, $N = 353$, $P < 0.001$. **b**, Zooplankton biomass versus phytoplankton biomass. English Channel data: $\log[y] = -0.93 + 1.96\log[x] - 0.37(\log[x])^2$; $R^2 = 0.51$, $N = 353$, $P < 0.001$. **c**, Microzooplankton diversity versus phytoplankton diversity. Global data set; diversity expressed by Shannon–Weaver diversity index³; Pearson correlation: $y = 1.38 + 0.10x$; $R^2 = 0.01$, $N = 353$, $P < 0.05$. **d**, Zooplankton diversity versus phytoplankton diversity. English Channel data: $y = 1.91 + 0.09x$; $R^2 < 0.01$; $N = 353$, n.s.). Data in **a** and **c** are from the Norwegian Sea (white uptriangles), the Iceland Basin (black uptriangles), Irminger Sea (white circles), Long Island Sound (black squares), North Sea (black diamonds), Benguela Upwelling (black downtriangles), Oregon upwelling (crosses), Indian Ocean (white diamonds), mesocosms in the Bergen fjord (white squares), and meridional transects in the Atlantic Ocean (black circles).

biomass (Fig. 3a). Several species of small nanophytoplankton (global data set) and picophytoplankton species (AMT4 data set; see Supplementary Information) were typically co-dominant at low phytoplankton biomass, consistent with earlier studies¹⁴. Large phytoplankton species dominated at high phytoplankton biomass. Phytoplankton communities with a very high phytoplankton biomass were almost invariably dominated by a single phytoplankton species (Fig. 3b). Close inspection of the global data set revealed that the dominant species at high phytoplankton biomass was generally

a diatom, a dinoflagellate, a coccolithophorid (*Emiliania huxleyii*) or a *Phaeocystis* species.

In view of the broad correlation between phytoplankton biomass and marine primary production¹⁵, we interpret the unimodal relation in Fig. 2 as a classic unimodal diversity–productivity pattern. A number of factors may be responsible for the unimodal

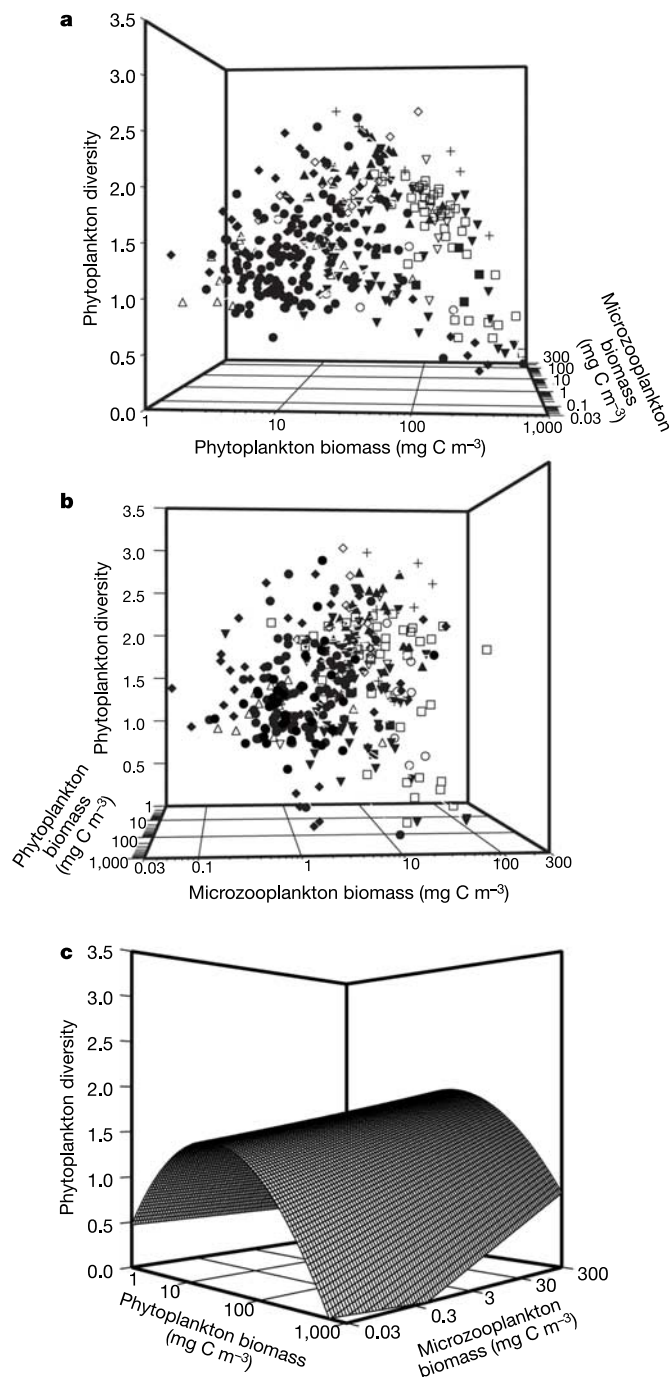


Figure 2 Biodiversity patterns of marine phytoplankton (global data set). **a**, Phytoplankton diversity, expressed by the Shannon–Weaver diversity index³, plotted as a function of phytoplankton biomass and microzooplankton biomass. **b**, The same data viewed from a different angle. **c**, Response surface calculated by multiple regression (Table 1). Symbols are as in Fig. 1.

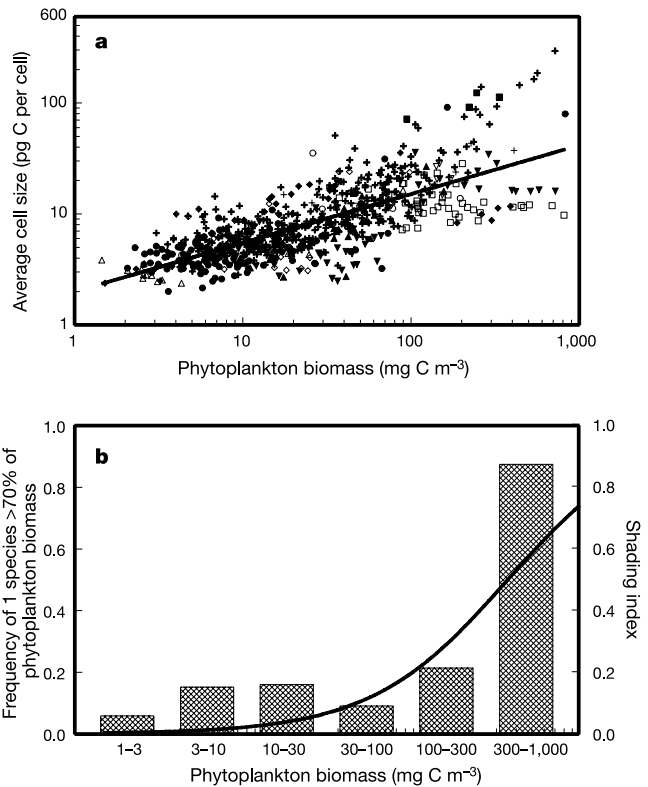


Figure 3 Specific patterns along the productivity gradient. **a**, Average phytoplankton cell size versus phytoplankton biomass. Nano- and microphytoplankton from all data sets: linear regression: $\log[y] = 0.30 + 0.44\log[x]$; $R^2 = 0.55$, $N = 720$, $P < 0.001$. Symbols are as in Fig. 1, with data from the English Channel indicated by bold crosses. **b**, Relation between species dominance and shading (nano- and microphytoplankton from all data sets). Vertical bars indicate the frequency of samples in which a single phytoplankton species made up more than 70% of the phytoplankton biomass. The solid line indicates the shading index as a function of phytoplankton biomass.

Table 1 Multiple regression analysis of plankton diversity versus plankton biomass

	Global data set		English Channel data	
	Value $\pm$ s.e.m.	P	Value $\pm$ s.e.m.	P
Phytoplankton diversity				
Constant	0.470 $\pm$ 0.132	<0.001	-0.383 $\pm$ 0.296	n.s.
$\log[\text{PB}]$	1.662 $\pm$ 0.195	<0.001	2.948 $\pm$ 0.397	<0.001
$(\log[\text{PB}])^2$	-0.597 $\pm$ 0.063	<0.001	-1.149 $\pm$ 0.139	<0.001
$\log[\text{ZB}]$	n.s.	n.s.	-0.744 $\pm$ 0.250	<0.005
$(\log[\text{ZB}])^2$	n.s.	n.s.	n.s.	n.s.
$\log[\text{PB}] \times \log[\text{ZB}]$	0.099 $\pm$ 0.039	<0.05	0.613 $\pm$ 0.151	<0.001
Overall significance	$R^2 = 0.21$	<0.001	$R^2 = 0.22$	<0.001
Zooplankton diversity				
Constant	1.423 $\pm$ 0.136	<0.001	1.764 $\pm$ 0.098	<0.001
$\log[\text{PB}]$	0.473 $\pm$ 0.194	<0.05	0.219 $\pm$ 0.064	<0.005
$(\log[\text{PB}])^2$	-0.267 $\pm$ 0.062	<0.001	n.s.	n.s.
$\log[\text{ZB}]$	n.s.	n.s.	0.581 $\pm$ 0.134	<0.001
$(\log[\text{ZB}])^2$	-0.630 $\pm$ 0.087	<0.001	-0.472 $\pm$ 0.060	<0.001
$\log[\text{PB}] \times \log[\text{ZB}]$	0.477 $\pm$ 0.081	<0.001	n.s.	n.s.
Overall significance	$R^2 = 0.16$	<0.001	$R^2 = 0.23$	<0.001

PB = phytoplankton biomass (mg C m⁻³); ZB = zooplankton biomass (mg C m⁻³). The global data set includes microzooplankton only ($N = 353$). The English Channel data include microzooplankton and mesozooplankton ($N = 353$). Terms that were not significant (n.s.) at the 0.05 level were removed from the regression equation.

diversity–productivity relationship^{2–5}. Traditionally, the species composition of marine phytoplankton has been explained in terms of nutrient availability and turbulence¹⁶. Small cells have a surface-to-volume ratio that is more favourable for acquiring nutrients at low nutrient concentrations¹⁷. Consequently, we conjecture that nutrient limitation is responsible for the co-dominance of a limited number of small phytoplankton species at low phytoplankton biomass.

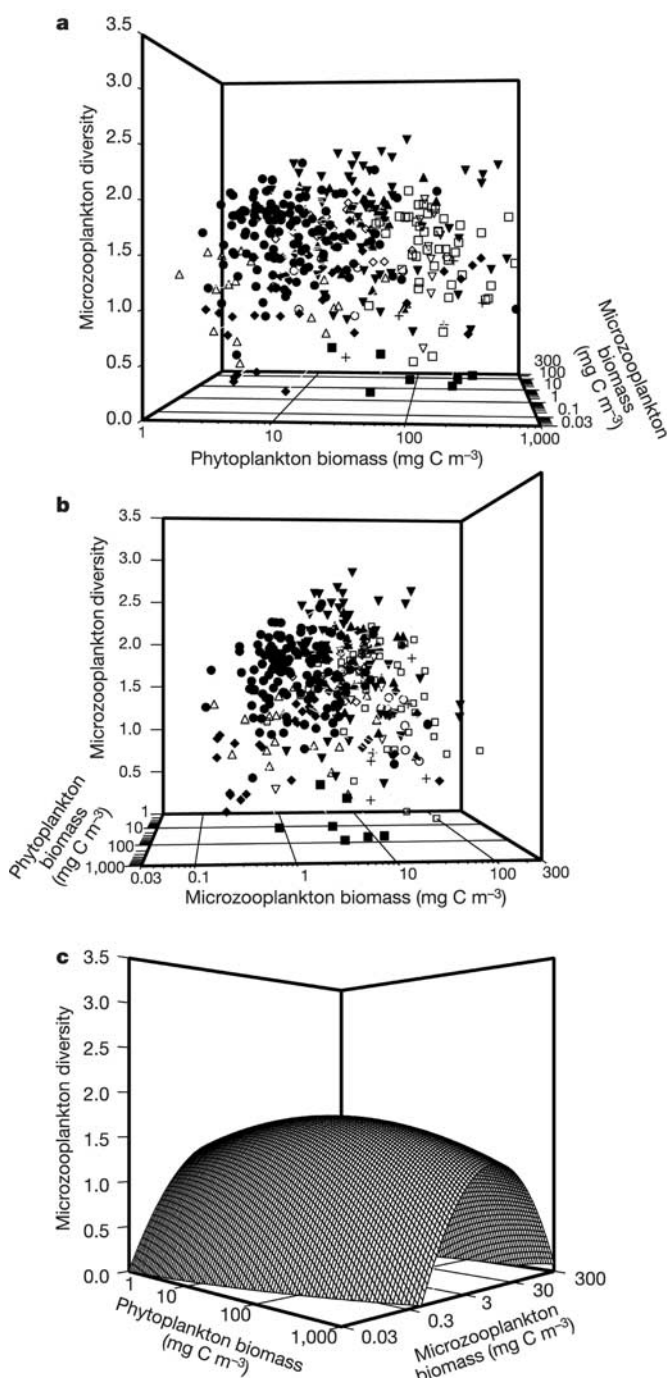


Figure 4 Biodiversity patterns of marine microzooplankton. **a**, Microzooplankton diversity, expressed by the Shannon–Weaver diversity index³, plotted as a function of phytoplankton biomass and microzooplankton biomass. **b**, The same data viewed from a different angle. **c**, Response surface calculated by multiple regression (Table 1). Symbols are as in Fig. 1.

In our samples, only large or protected phytoplankton species (dinoflagellates, *Phaeocystis* colonies, diatoms, coccolithophorids) that can escape microzooplankton predation because of their size, chemical or mechanical defences^{18,19} were able to bloom at a phytoplankton biomass higher than 100 mg m⁻³. This is consistent with the observation in Fig. 1a that grazing is less intense at high phytoplankton biomass. Recent studies indicate that competition for light^{20,21} could also explain the low diversity during intense phytoplankton blooms.

To test this hypothesis, we used a shading index, S , that calculates the fraction of the total light attenuation in the water column that is due to light attenuation by phytoplankton, according to $S = kC/(kC + K_{bg})$. Here, k is the specific light attenuation coefficient of chlorophyll a , C is the chlorophyll a concentration, and K_{bg} is the background light attenuation of the water column. Literature data²² indicate that, as a rough approximation, $k \approx 0.014 \text{ m}^2 (\text{mg Chl } a)^{-1}$, the carbon to chlorophyll a ratio of marine phytoplankton is ~ 50 , and $K_{bg} \approx 0.10 \text{ m}^{-1}$. A plot of the shading index against phytoplankton biomass reveals that phytoplankton diversity declined when shading within the phytoplankton community increased (Fig. 3b). We therefore conjecture that a combination of selective grazing and shading is responsible for the low phytoplankton diversity at high phytoplankton biomass.

Previous work has argued that nutrients and herbivores have interactive effects on the diversity of primary producers^{23–25}. That is, at higher levels of zooplankton biomass the phytoplankton diversity peaks at a higher productivity. We found only a weak interaction effect of phytoplankton biomass and microzooplankton biomass on phytoplankton diversity (Table 1, Fig. 2). A stronger interaction effect was found when mesozooplankton was included (Table 1), suggesting a shift in the predation source from microzooplankton to mesozooplankton with increasing phytoplankton biomass.

Zooplankton diversity is a unimodal function of zooplankton biomass, both in the global data set with microzooplankton (Fig. 4) and in the English Channel data including mesozooplankton (see Supplementary Information). We speculate that the unimodal pattern in zooplankton diversity can be interpreted similarly to the unimodal pattern in phytoplankton diversity, as a balance between food limitation at low population levels and selective predation at high population levels.

Our data set, from 12 different marine areas, confirms plankton diversity patterns reported earlier for specific aquatic regions^{8,9,11,12}. The worldwide consistency of the plankton patterns documented here is striking, since the various oceanographic regions sampled have very different nutrient, light and turbulence conditions²⁶, and the physiologies of the various plankton groups are also very different. Moreover, the diversity–productivity relation in oceanic plankton documented here is similar to the classic diversity–productivity relation observed in terrestrial systems^{1–3}, both in terms of the unimodal patterns and in terms of the explanations involved. We thus conclude that universal mechanisms underlie the diversity patterns of plants and herbivores in both terrestrial and marine habitats. □

Methods

Water samples for species identification and carbon estimation of phytoplankton and microzooplankton were collected in the Norwegian Sea ($N = 19$), the North Atlantic Ocean ($N = 8$), the Iceland Basin ($N = 25$), the Irminger Sea ($N = 8$), Long Island Sound ($N = 7$), the North Sea ($N = 41$), the Benguela upwelling ($N = 51$), the Oregon upwelling ($N = 7$), the Indian Ocean ($N = 16$), mesocosms in the Bergen fjord ($N = 45$), and five AMTs from 48°N to 50°S in the Atlantic Ocean ($N = 126$). When several depths were sampled on the same cast only the depth with maximum phytoplankton concentration was considered for this analysis. This global data set consisted of 353 water samples. For the picophytoplankton data on AMT4 all depths where picophytoplankton and other phytoplankton were sampled simultaneously are included in the analysis ($N = 40$). Geographic coordinates, sampling dates, and depths of the sampling stations can be found in the additional material. For the English Channel we used a separate database of 353 water samples for analyses of phytoplankton, microzooplankton and mesozooplankton. Sampling details and data of the latter are available at <http://www.pml.ac.uk/L4>. Two

replicate samples were preserved, one with 1% final concentration Lugol's iodine solution and the other with 1% buffered formalin²⁷. Subsamples (100 ml) were settled (Utermöhl technique) and counted at the species level with an inverted microscope. Biomass is expressed in terms of carbon. Phytoplankton and heterotrophic dinoflagellates carbon biomass was estimated from known relationships with cell volume²⁸. Ciliate carbon biomass was estimated using a factor of 0.21 pg C μm^{-3} (ref. 29). Copepod biomass was calculated using species-specific conversion factors³⁰.

Received 10 March; accepted 22 April 2004; doi:10.1038/nature02593.

1. Grime, J. P. Competitive exclusion in herbaceous vegetation. *Nature* **242**, 344–347 (1973).
2. Tilman, D. *Resource Competition and Community Structure* (Princeton Univ. Press, Princeton, 1982).
3. Rosenzweig, M. L. *Species Diversity in Space and Time* (Cambridge Univ. Press, Cambridge, 1995).
4. Waide, R. B. *et al.* The relationship between productivity and species richness. *Annu. Rev. Ecol. Syst.* **30**, 257–300 (1999).
5. Mittelbach, G. G. *et al.* What is the observed relationship between species richness and productivity? *Ecology* **82**, 2381–2396 (2001).
6. Chase, J. M. & Leibold, M. A. Spatial scale dictates the productivity–biodiversity relationship. *Nature* **416**, 427–430 (2002).
7. Fukami, T. & Morin, P. J. Productivity–biodiversity relationships depend on the history of community assembly. *Nature* **424**, 423–426 (2003).
8. Williams, L. Possible relationships between plankton–diatom species numbers and water-quality estimates. *Ecology* **45**, 809–823 (1964).
9. Dodson, S. L., Arnott, S. E. & Cottingham, K. L. The relationship in lake communities between primary productivity and species richness. *Ecology* **81**, 2662–2679 (2000).
10. Kassen, R., Buckling, A., Bell, G. & Rainey, P. B. Diversity peaks at intermediate productivity in a laboratory microcosm. *Nature* **406**, 508–512 (2000).
11. Interlandi, S. J. & Kilham, S. S. Limiting resources and the regulation of diversity in phytoplankton communities. *Ecology* **82**, 1270–1282 (2001).
12. Li, W. K. W. Macroecological patterns of phytoplankton in the northwestern North Atlantic Ocean. *Nature* **419**, 154–157 (2002).
13. Calbet, A. & Landry, M. R. Phytoplankton growth, microzooplankton grazing, and carbon cycling in marine systems. *Limnol. Oceanogr.* **49**, 51–57 (2004).
14. Agawin, N. S. R., Duarte, C. M. & Agustí, S. Nutrient and temperature control of the contribution of picoplankton to phytoplankton biomass and production. *Limnol. Oceanogr.* **45**, 591–600 (2000).
15. Joint, I. & Groom, S. B. Estimation of phytoplankton production from space: current status and future potential of satellite remote sensing. *J. Exp. Mar. Biol. Ecol.* **250**, 233–255 (2000).
16. Margalef, R. Life-forms of phytoplankton as survival alternatives in an unstable environment. *Oceanol. Acta* **1**, 493–509 (1978).
17. Raven, J. A. Small is beautiful: the picophytoplankton. *Funct. Ecol.* **12**, 503–513 (1998).
18. Strom, S. *et al.* Chemical defense in the microplankton. I. Feeding and growth rates of the heterotrophic protists on the DMS-producing phytoplankter *E. huxleyi*. *Limnol. Oceanogr.* **48**, 217–229 (2003).
19. Hamm, C. E. *et al.* Architecture and material properties of diatom shells provide effective mechanical protection. *Nature* **421**, 841–843 (2003).
20. Huisman, J., Jonker, R. R., Zonneveld, C. & Weissing, F. J. Competition for light between phytoplankton species: experimental tests of mechanistic theory. *Ecology* **80**, 211–222 (1999).
21. Huisman, J., van Oostveen, P. & Weissing, F. J. Species dynamics in phytoplankton blooms: incomplete mixing and competition for light. *Am. Nat.* **154**, 46–68 (1999).
22. Kirk, J. T. O. *Light and Photosynthesis in Aquatic Ecosystems*, 2nd edn (Cambridge Univ. Press, Cambridge, 1994).
23. Collins, S. L., Knapp, A. K., Briggs, J. M., Blair, J. M. & Steinauer, E. M. Modulation of diversity by grazing and mowing in native tallgrass prairie. *Science* **280**, 745–747 (1998).
24. Worm, B., Lotze, H. K., Hillebrand, H. & Sommer, U. Consumer versus resource control of species diversity and ecosystem functioning. *Nature* **417**, 848–851 (2002).
25. Hillebrand, H. Opposing effects of grazing and nutrients on diversity. *Oikos* **100**, 592–600 (2003).
26. Longhurst, A. R. *Ecological Geography of the Sea* (Academic, London, 1998).
27. Holligan, P. M. & Harbour, D. S. The vertical distribution and succession of phytoplankton in the western English Channel in 1975 and 1976 (1977). *J. Mar. Biol. Assoc. UK* **57**, 1075–1093 (1977).
28. Strathmann, R. R. Estimating the organic carbon content of phytoplankton from cell volume or plasma volume. *Limnol. Oceanogr.* **12**, 411–418 (1967).
29. Ohman, M. D. & Runge, J. A. Sustained fecundity when phytoplankton resources are in short supply: omnivory by *Calanus finmarchicus* in the Gulf of St Lawrence. *Limnol. Oceanogr.* **39**, 21–36 (1994).
30. Mauchline, J. *The Biology of Calanoid Copepods* (Academic, San Diego, 1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank all who contributed to collecting the samples on the different cruises. Special thanks go to D. Harbour, who counted most of the samples to the species level and to M. Zubkov for the picoplankton data. X.I. was supported by a Ramon y Cajal grant from the Spanish Ministry for Science and Technology and the Departments of Agriculture, Fisheries and Education, and Universities and Research of the Basque Country Government. The research of J.H. was supported by the Earth and Life Sciences Foundation (ALW), which is subsidized by the Netherlands Organization for Scientific Research (NWO). The research of R.P.H. is a contribution to the Plymouth Marine Laboratory Core Strategic Research Programme. This study was supported by the UK Natural Environment Research Council through the Atlantic Meridional Transect consortium (this is contribution number 87 of the AMT programme).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to X.I. (xirigoien@pas.azti.es).

Coral communities are regionally enriched along an oceanic biodiversity gradient

Ronald H. Karlson¹, Howard V. Cornell¹ & Terence P. Hughes²

¹Department of Biological Sciences, University of Delaware, Newark, Delaware 19716, USA

²Centre for Coral Reef Biodiversity, Department of Marine Biology, James Cook University, Townsville, Queensland 4811, Australia

Ecological communities are influenced by processes operating at multiple scales^{1–5}. Thus, a better understanding of how broad- as well as local-scale processes affect species diversity and richness is increasingly becoming a central focus in modern community ecology^{6–9}. Here, in a study of unprecedented geographical scope, we show significant regional and local variation in the species richness of coral assemblages across an oceanic biodiversity gradient. The gradient that we sampled extends 10,000 km eastwards from the world's richest coral biodiversity hotspot in the central Indo-Pacific¹⁰. Local richness and the size of regional species pools decline significantly across 15 islands spanning the gradient. In addition, richness declines across three adjacent habitats (reef slopes, crests and flats). In each habitat, a highly consistent linear relationship between local and regional species richness indicates strong regional enrichment. Thus, even on the most diverse coral reefs in the world, local coral assemblages are profoundly affected by regional-scale processes. Understanding these historical and biogeographical influences is essential for the effective management and preservation of these endangered communities.

Strong evidence for regional enrichment of biodiversity is indicated by a linear relationship between local and regional species richness across a regional diversity gradient^{6,7}. Conversely, if local richness reaches an upper limit (that is, becomes independent of regional richness), then strong regional enrichment is not supported. Until now, evidence for regional enrichment of coral assemblages was based on meta-analyses of 63 disparate studies that showed only weak support for regional enrichment^{8,11}. This earlier work established that local coral assemblages in species-rich regions are more diverse than those in less species-rich regions, but did not permit discrimination between two models representing the theoretical extremes for such relationships (linear type I versus levelling type II⁷). The putative relationship between local and regional richness was curvilinear and confounded by the prevalence of small samples with relatively few coral colonies¹². Furthermore, the strongest evidence for regional enrichment arose only from depauperate regions of the Indo-Pacific. These preliminary results may have been influenced by biases in the biogeographical location of published studies (for example, inadequate sampling near the central Indo-Pacific biodiversity hotspot), and by the inherent variability of local richness estimates (owing to methodological differences among studies and the wide range of depths and habitats that were investigated). Here, we overcome these limitations using standardized hierarchical sampling of local richness in each of three habitats along an oceanic-scale richness gradient emanating from a major biodiversity hotspot (Fig. 1). It is in such species-rich hotspots where theory predicts that local richness is most likely to be limited^{13,14}, and where regional enrichment is least likely to occur¹².

Local species richness varies substantially among the three habitats and five geographical regions over the richness gradient (Fig. 2). For each habitat, hierarchical analysis of variance (ANOVA)¹⁵ on the number of species observed per site indicates very strong regional differences ($F = 19.67, 49.68$ and 15.59 for reef

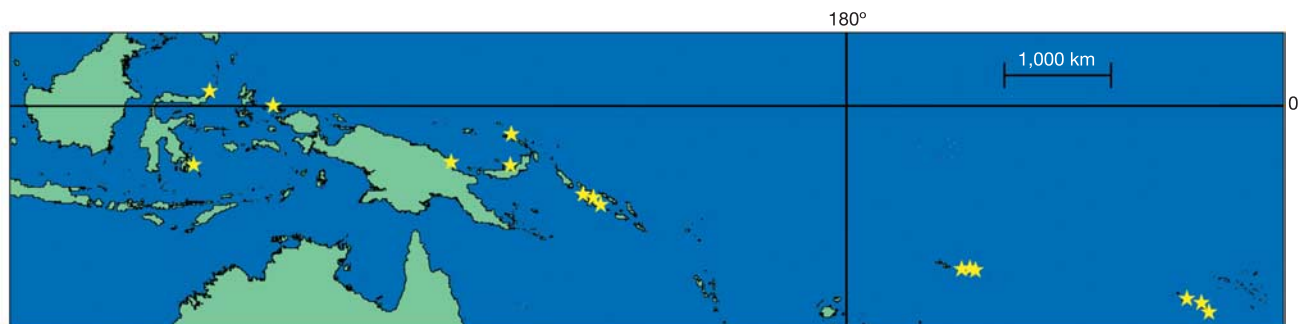


Figure 1 Fifteen island locations distributed across five Indo-Pacific regions. Stars correspond with the following localities: Irian Jaya, Northern and Southern Sulawesi in Indonesia; Madang, New Britain and New Ireland in Papua New Guinea; Gizo, Munda, and

Upei Island in the Solomon Islands; Tutuila, Tau and Ofu-Olosega in American Samoa; and Raiatea, Moorea and Tahiti in the Society Islands of French Polynesia.

slopes, crests and flats, respectively, $P < 0.001$). In contrast, the differences among islands within regions are either very weak ($F = 2.15$ for reef slopes, $P < 0.05$) or insignificant ($F = 1.17$ and 1.75 for reef crests and flats, respectively, $P > 0.05$). Local richness in western regions is higher than in the central Pacific in all three habitats, although the pattern of change varies among habitats. On reef slopes, richness is high in Indonesia, Papua New Guinea and the Solomon Islands but sharply lower further east in American Samoa and the Society Islands (Fig. 2). On reef crests, richness drops sharply between Indonesia and Papua New Guinea and then decreases more gradually eastwards. On reef flats, richness also declines gradually towards the east but with fewer significant differences among regions than on reef crests or slopes. Local species richness is highest on reef slopes, intermediate on crests, and lowest on reef flats almost everywhere.

Two other related measures of local diversity (Fisher's alpha¹⁶ and the estimated number of species in each habitat at each site¹⁷) yield very similar results to those depicted in Fig. 2. Fisher's alpha is widely recognized as a stable, relatively sample-size-independent diversity measure at both local and regional scales¹⁸, and both measures are often used to compensate for undersampling. Fisher's alpha is most appropriate when applied to logarithmic series distributions representing samples with large numbers of rare species and few common ones. In each habitat, highly significant regional variation is observed for each of these diversity measures

($F \geq 13.01$, $P < 0.001$) whereas the among-island variation again is quite weak (Fisher's alpha: $F = 2.19$ for reef slopes, $P < 0.05$, $F \leq 1.74$ for reef crests and flats, $P > 0.05$; 'Chao 1' estimator (see Methods): $F \leq 1.62$ for reef slopes, crests and flats, $P > 0.05$). The consistency of these results indicates that undersampling is not a problem in this analysis.

The total species richness estimated using a standardized effort of 120 transects per region also varies considerably among regions and habitats (Fig. 3). In general, regional richness for reef flats, crests and slopes decreases from Indonesia eastward and increases from reef flats towards reef slopes (Fig. 3). Regional richness is highest on reef slopes in Indonesia (235 species) and lowest on reef flats in American Samoa (37 species). The regional species pool for reef slopes is twice the size of that for flats and approximately 1.5 times that for crests. Comparable values of Fisher's alpha are virtually identical with the regional diversity for reef slopes being 1.9 times that for flats and 1.5 times that for crests. These estimates of the relative sizes of the regional species pools are based on the

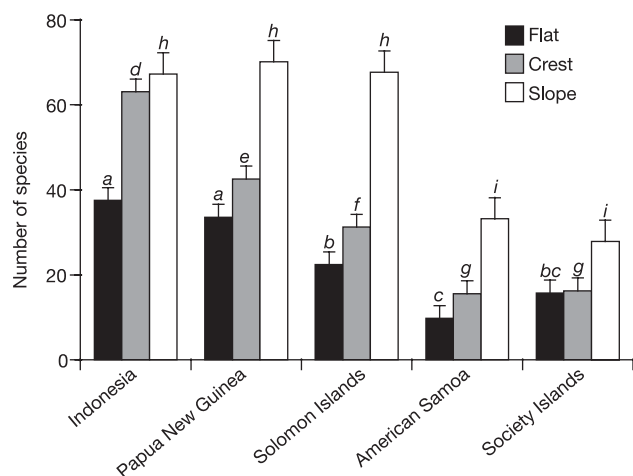


Figure 2 Local species richness across regions and habitats. The mean local species richness per site (± 1 standard error, $n = 12$ sites distributed across three islands per region) is plotted for each of three habitats across five regions along the Pacific diversity gradient (Fig. 1). Letters designate significantly different means for reef flats (a–c), crests (d–g) and slopes (h, i) as determined using Duncan's multiple range tests ($P < 0.05$).

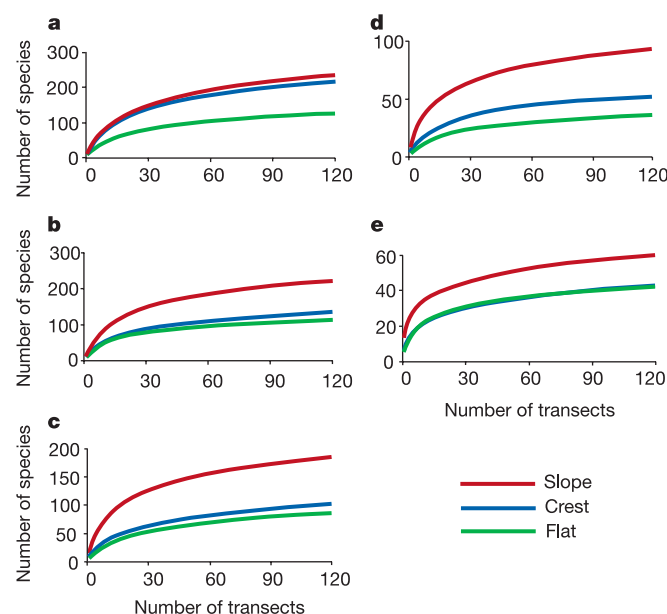


Figure 3 Observed species accumulation curves across regions and habitats. **a–e**, The cumulative number of species is plotted for 120 transects within each of five regions (Indonesia (**a**), Papua New Guinea (**b**), the Solomon Islands (**c**), American Samoa (**d**) and the Society Islands (**e**)) and three habitats (reef flats, crests and slopes). Each curve depicts the average of 100 replicate randomizations of transect order. The total number of species observed in 120 transects is used as a standardized measure of regional species richness in each habitat.

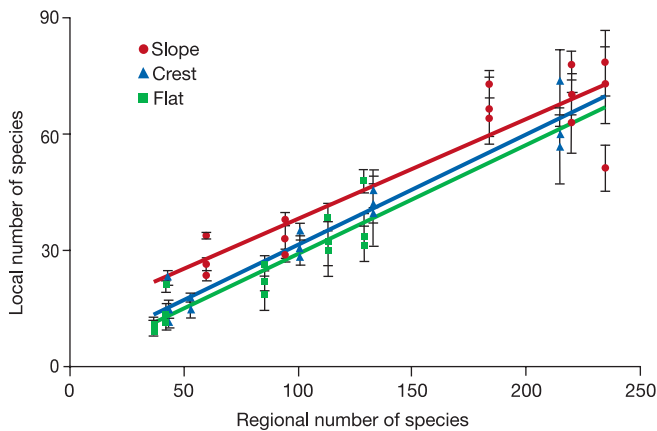


Figure 4 The relationship between local and regional species richness in corals. Local richness is plotted as the mean number of species per site (± 1 standard error, $n = 4$ sites per island) on reef flats (squares), crests (triangles) and slopes (circles). Regional richness is the cumulative total number of species sampled in 120, 10-m transects per region (Fig. 3). Linear regressions are presented for the mean local richness per site in a region (y) versus regional richness (x): flats, $y = 1.15 + 0.28x$, standard error of slope = 0.03, $R^2 = 0.97$; crests, $y = 2.72 + 0.28x$, standard error of slope = 0.02, $R^2 = 0.99$; slopes, $y = 11.98 + 0.26x$, standard error of slope = 0.04, $R^2 = 0.94$.

occurrence of coral colonies within each of the specific habitats we sampled. Estimates of the absolute sizes of these regional pools would be higher if one were to pool occurrences across all coral habitats (as has been done previously^{8,11,12}), but this would ignore the observed twofold difference in regional diversity between reef flats and slopes. Over all five regions, we encountered a total of 333 species in three reefal habitats. This number represents more than half of all coral species known to occur along this Pacific diversity gradient¹⁰ across a much more diverse range of depths and habitats. Because our samples exclude coral species restricted to additional habitats (such as deep reef slopes and lagoons), we evaluate richness using habitat-specific data.

Regional enrichment is strongly supported across the diversity gradient as indicated by a remarkably consistent linear type I relationship between local and regional habitat-specific richness (Fig. 4). The slopes of the linear regressions are virtually identical, indicating that the proportion of species sampled within each habitat at each site is 27% of the regional pool. Comparable regressions using Fisher's alpha, the Chao 1 estimator, and numbers of species in samples standardized to 100 or more colonies¹⁸ are also linear. There is little apparent variation in the number of species per site among islands within regions, except perhaps on Indonesian slopes (Fig. 4) where the relatively low observed local richness in Irian Jaya (51 species per site) contributes to a weak but significant difference among islands. However, this difference is not significant when local richness is estimated using Chao 1. In contrast, variation due to regional differences is highly significant regardless of the diversity measure used.

There are several potential mechanisms responsible for the strong regional enrichment of local coral assemblages^{7,12}. Enriched assemblages are open to the influence of historical and geographical phenomena characterizing regions (for example, oceanic transport, location relative to biodiversity hotspots, and evolutionary processes generating variation in the size of regional species pools). These influences penetrate local habitats because spatial heterogeneity and species coexistence are promoted by disturbances, spatial aggregation, or biological interactions. Even when the potential for competitive exclusion is intense, theory predicts that positive relationships between local and regional richness are likely under a wide range of conditions provided that immigration and/or local extinction rates are sufficiently high^{19,20} or that natural enemies

behave like keystone predators²¹. Most rigorous tests of hypothetical mechanisms promoting habitat-wide species coexistence (such as those invoking competitive neighbourhoods or intraspecific aggregation) have been performed in terrestrial plant communities^{22,23}. With the demonstration of significant regional enrichment of coral assemblages in the Indo-Pacific, tests are now needed to discriminate between the alternative mechanistic explanations for species coexistence in corals.

Our results have profound implications for understanding the scale-dependent structure and dynamics of coral assemblages. Local coral assemblages are embedded in regions with different histories, geographies and human influences, and they are influenced jointly by both local and regional factors. Thus a blend of ecological and biogeographical studies at multiple scales are needed to fully understand coral biodiversity. Recent assessments of the status of coral reefs verify that they are globally threatened and efforts to manage them will require international cooperation²⁴. As most conservation efforts by policy makers and managers begin with attempts to preserve local coral reefs, it is imperative that we learn how coral communities function at these local, ecologically important scales while maintaining a regional perspective. Our results document the importance of this regional approach, yet much remains to be learned about the dynamics of coral reefs. We need to focus our studies on how local populations are maintained by stock-recruitment relationships and immigration from both local and regional sources. Similarly, the dynamics of communities must be examined at local and metacommunity scales^{25,26}. Understanding such scale-dependent processes will be essential if we are to be successful at managing and preserving these important biological structures.

In this context, there are several relevant results from our study. First, we demonstrate that the diversity of local coral assemblages is strongly related to regional differences in species richness (Fig. 4). Thus, local diversity patterns are influenced by more than local environmental factors and a broader regional perspective is required to understand them. Second, significant regional variation in local diversity extends along the entire 10,000-km biodiversity gradient we sampled. Thus, regional enrichment is pervasive and a relevant factor influencing both species-rich and relatively depauperate assemblages (Fig. 4). Third, regional species richness varies substantially between major coral reef habitats. This study is the first to quantify this difference and to make ecologically relevant predictions regarding the relative size of regional species pools for corals on reef flats, crests and slopes. Clearly, the dynamics of coral assemblages must be studied at multiple scales. For example, ecological processes operating at the scale of individual transects, local sites and larger metacommunities represent potentially important sources of variation in local diversity on coral reefs. Fourth, the habitat-specific evaluation of regional enrichment indicates a consistent proportional relationship between local and regional species richness common to this set of three reefal habitats. This relationship may apply to other habitats for which few data are currently available. □

Methods

Sampling

We partitioned samples among three common habitats, matched among regions: reef flats, crests and slopes. Reef flats were approximately 5–10 m inshore of breaking waves, whereas reef crests and slopes were seaward at depths of 1–2 m and 6–7 m, respectively. There were four replicate sites per island and three islands in each of five regions (Fig. 1). Over all regions and habitats, we sampled a total of 41,710 coral colonies along 1,800 10-m line transects. At each site, ten 10-m line transects were laid parallel to depth contours, draped over the substratum in each habitat, and all coral colonies intercepted by the tape were identified to species level. Training in coral identification was provided by the Museum of Tropical Queensland, which houses the most extensive scleractinian collection in the world. Four key publications^{27–30} were particularly useful during this training.

Statistical analysis

Measures of local diversity (species richness, Fisher's alpha, and the estimated species richness per site using the Chao 1 procedure) were evaluated for significant variation

among islands and regions using nested, mixed-model ANOVA. We screened several potential estimators to find that the Chao 1 procedure provided the most stable values for local species richness. This estimator is the sum of the observed number of species and the quotient $a^2/2b$, where a and b equal the number of species represented by one and two colonies, respectively.

To analyse the local–regional species richness relationship in each habitat, we used a simple linear regression of the mean local richness per site in a region against habitat-specific regional richness. Linearity is supported by these and supplemental regressions using: (1) log-transformed richness data; (2) local richness standardized to 100 or more colonies per sample¹⁸; and (3) the two alternative measures of local species diversity, Fisher's alpha and the Chao 1 estimator of local richness.

Received 14 November 2003; accepted 24 May 2004; doi:10.1038/nature02685.

Published online 16 June 2004.

1. Ricklefs, R. E. & Schluter, D. (eds) *Species Diversity in Ecological Communities: Historical and Geographical Perspectives* (Univ. Chicago Press, Chicago, Illinois, 1993).
2. Levin, S. A. The problem of pattern and scale in ecology. *Ecology* **73**, 1943–1967 (1992).
3. Roughgarden, J., May, R. M. & Levin, S. A. (eds) *Perspectives in Ecological Theory* (Princeton Univ. Press, Princeton, New Jersey, 1989).
4. Yodanis, P. *Introduction to Theoretical Ecology* (Harper & Row, New York, 1989).
5. Tilman, D. & Kareiva, P. (eds) *Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions* (Princeton Univ. Press, New Jersey, 1997).
6. Ricklefs, R. E. Community diversity: relative roles of local and regional processes. *Science* **235**, 167–171 (1987).
7. Cornell, H. V. & Lawton, J. H. Species interactions, local and regional processes, and limits to the richness of ecological communities: a theoretical perspective. *J. Anim. Ecol.* **61**, 1–12 (1992).
8. Cornell, H. V. & Karlson, R. H. Diversity of reef-building corals determined by local and regional processes. *J. Anim. Ecol.* **65**, 233–241 (1996).
9. Caley, M. J. & Schluter, D. The relationship between local and regional diversity. *Ecology* **78**, 70–80 (1997).
10. Hughes, T. P., Bellwood, D. R. & Connolly, S. R. Biodiversity hotspots, centres of endemism, and the conservation of coral reefs. *Ecol. Lett.* **5**, 775–784 (2002).
11. Karlson, R. H. & Cornell, H. V. Scale-dependent variation in local vs. regional effects on coral species richness. *Ecol. Monogr.* **68**, 259–274 (1998).
12. Karlson, R. H. & Cornell, H. V. Species richness of coral assemblages: detecting regional influences at local scales. *Ecology* **83**, 452–463 (2002).
13. Elton, C. *The Ecology of Animals* (Methuen, London, 1933).
14. MacArthur, R. H. Patterns of species diversity. *Biol. Rev.* **40**, 510–533 (1965).
15. Hughes, T. P. *et al.* Patterns of recruitment and abundance of corals along the Great Barrier reef. *Nature* **397**, 59–63 (1999).
16. Fisher, R. A., Corbet, A. S. & Williams, C. B. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**, 42–58 (1943).
17. Colwell, R. K. & Coddington, J. A. Estimating terrestrial biodiversity through extrapolation. *Phil. Trans. R. Soc. Lond. B* **345**, 101–118 (1994).
18. Hayek, L.-A. C. & Buzas, M. A. *Surveying Natural Populations* (Columbia Univ. Press, New York, 1997).
19. Caswell, H. & Cohen, J. E. in *Species Diversity in Ecological Communities: Historical and Geographical Perspectives* (eds Ricklefs, R. E. & Schluter, D.) 99–107 (Univ. Chicago Press, Chicago, Illinois, 1993).
20. Loreau, M. & Mouquet, N. Immigration and the maintenance of local species diversity. *Am. Nat.* **154**, 427–440 (1999).
21. Shurin, J. B. & Allen, E. G. Effects of competition, predation, and dispersal on species richness at local and regional scales. *Am. Nat.* **158**, 624–637 (2001).
22. Pacala, S. W. & Silander, J. A. Jr Field tests of neighborhood population dynamic models of two annual weed species. *Ecol. Monogr.* **60**, 113–134 (1990).
23. Stoll, P. & Prati, D. Intraspecific aggregation alters competitive interactions in experimental plant communities. *Ecology* **82**, 319–327 (2001).
24. Hughes, T. P. *et al.* Climate change, human impacts, and the resilience of coral reefs. *Science* **301**, 929–933 (2003).
25. Holt, R. D. in *Species Diversity in Ecological Communities: Historical and Geographical Perspectives* (eds Ricklefs, R. E. & Schluter, D.) 77–88 (Univ. Chicago Press, Chicago, Illinois, 1993).
26. Holt, R. D. in *Metapopulation Biology: Ecology, Genetics, and Evolution* (eds Hanski, I. A. & Gilpin, M. E.) 149–164 (Academic, San Diego, California, 1997).
27. Veron, J. E. N., Pichon, M. & Wijsman-Best, M. *Scleractinia of Eastern Australia*. Part II Families Faviidae, Trachyphyllidae: Australian Institute of Marine Science Monograph Series (Australian Government Publishing Service, Canberra, 1977).
28. Veron, J. E. N. & Pichon, M. *Scleractinia of Eastern Australia*. Part IV Family Poritidae: Australian Institute of Marine Science Monograph Series (Australian National University Press, Canberra, 1982).
29. Wallace, C. C. *Staghorn Corals of the World* (CSIRO, Collingwood, Victoria, 1999).
30. Veron, J. E. N. *Corals of the World* (Australian Institute of Marine Sciences, Townsville, Queensland, 2000).

Acknowledgements We acknowledge J. Wolstenholme and C. Wallace of the Museum of Tropical Queensland for taxonomic training. We also thank J. Wolstenholme, M. J. Boyle, L. Dinsdale, A. Green, A. Baird, S. Connolly, M. Kospartov, P. Osmond and K. Osmond for their assistance in the field. Funding was provided by grants from the Australian Research Council, the National Science Foundation (International Programs and the Ocean Sciences Division) and the National Geographic Society.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.H.K. (rkarlson@udel.edu).

Global patterns in human consumption of net primary production

Marc L. Imhoff¹, Lahouari Bounoua^{1,2}, Taylor Ricketts^{3,4}, Colby Loucks³, Robert Harriss⁵ & William T. Lawrence⁶

¹Biospheric Sciences Branch, NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

²Earth System Science Interdisciplinary Center, University of Maryland, College Park, Maryland 20742, USA

³World Wildlife Fund, 1250 Twenty-fourth Street NW, Washington DC 20037, USA

⁴Department of Biological Sciences, Stanford University, Stanford, California 94305, USA

⁵Environmental & Societal Impacts Group, National Center for Atmospheric Research, Foothills Laboratory, 3450 Mitchell Lane, Boulder, Colorado 80307-3000, USA

⁶Department of Natural Sciences, Bowie State University, Bowie, Maryland 20715, USA

The human population and its consumption profoundly affect the Earth's ecosystems^{1,2}. A particularly compelling measure of humanity's cumulative impact is the fraction of the planet's net primary production that we appropriate for our own use^{3,4}. Net primary production—the net amount of solar energy converted to plant organic matter through photosynthesis—can be measured in units of elemental carbon and represents the primary food energy source for the world's ecosystems. Human appropriation of net primary production, apart from leaving less for other species to use, alters the composition of the atmosphere⁵, levels of biodiversity⁶, energy flows within food webs⁷ and the provision of important ecosystem services⁸. Here we present a global map showing the amount of net primary production required by humans and compare it to the total amount generated on the landscape. We then derive a spatial balance sheet of net primary production 'supply' and 'demand' for the world. We show that human appropriation of net primary production varies spatially from almost zero to many times the local primary production. These analyses reveal the uneven footprint of human consumption and related environmental impacts, indicate the degree to which human populations depend on net primary production 'imports' and suggest policy options for slowing future growth of human appropriation of net primary production.

An influential study³ has estimated global human appropriation of terrestrial net primary production (HANPP) to be 31% of the total amount of net primary production (NPP) generated on land, with 'low' and 'high' estimates of 3% and 39%, respectively. A more recent study, using the same definition of appropriation, estimated a similar value for global HANPP (32%)⁴, but high uncertainty was reported (10–55%; see ref. 7). This is a remarkable level of co-option for a species that represents roughly 0.5% of the total heterotroph biomass on Earth⁹. These previous estimates, however, were derived from global and biome averages, so that important spatial patterns in both NPP and HANPP remain hidden^{7,10}. Understanding these patterns is critical for identifying areas experiencing severe human impacts on their ecosystems, for illuminating the effects of different consumption patterns among regions and cultures, and for indicating the directions of net energy flow due to regional and global trade.

To examine these issues we developed a global map of HANPP, which allows spatial comparison to NPP within the context of the global carbon cycle. We defined HANPP as the amount of terrestrial NPP required to derive food and fibre products consumed by

humans, including the organic matter that is lost during the harvesting and processing of whole plants into end products. We started with data from the Food and Agriculture Organization (FAO) on products consumed in 1995, for 230 countries in seven categories: vegetal foods, meat, milk, eggs, wood (for building and fuel), paper and fibre¹¹. To these data we applied harvest, processing and efficiency multipliers, as well as estimates of below-ground production, to reconstruct the total amount of NPP required to derive the final products (Table 1, see Methods). We then calculated the per capita HANPP of each country and applied these values to a gridded database of the human population with a 0.25° spatial

resolution¹² (this resolution was chosen to match that of the NPP data, see below). Our method assumes a homogeneous per capita consumption rate within each country, which, although unsupported in some cases, is a reasonable starting point^{7,10}. In addition, terrestrial HANPP does not directly capture other important forms of environmental impact, including freshwater co-option, use of fossil fuels and appropriation of NPP from freshwater or marine systems¹.

The resulting map depicts the spatial pattern of HANPP, showing where the products of terrestrial photosynthesis are consumed (Fig. 1a). Summing for the globe, we estimated annual HANPP to

Table 1 Product multipliers for estimation of HANPP

Product/multiplier*	High estimate	Intermediate estimate	Low estimate
Vegetal food, fibre and grain fed to livestock†			
Post harvest losses ^{20,21}	0.4	0.10 (I) 0.40 (D)	0.1
Crop residue ²²	1.28	0.71	0.14
Milk efficiency factor ^{26,27} (kg grain/kg milk)	0.3	0.3 (I) 0.20 (D)	0.2
Eggs efficiency factor ^{26,27} (kg grain/kg eggs)	2.2	2.2 (I) 1.6 (D)	1.6
Wood and paper			
Milling loss ²³	0.34	0.07 (I) 0.34 (D)	0.07
Harvest loss ²⁴	0.54	0.22 (I) 0.54 (D)	0.22
Recycling (paper only) ²⁵	−0.20	−0.25 (I) −0.20 (D)	−0.25
Carbon/organic matter conversion ²⁹	0.50	0.475	0.45
Roots (short/tall vegetation) ¹³	2.0/1.5	2.0/1.5	2.0/1.5

* Multipliers were used to cumulatively add organic matter to the mass of FAO-reported products in a logical sequence (from top down in the table). Multipliers were either applied equally to all countries or were applied separately to industrialized (I) and developing countries (D). Development status was assigned according to ref. 22.

† For meat products, the only variable adjusted to derive high and low estimates is the efficiency with which grain feeds are produced.

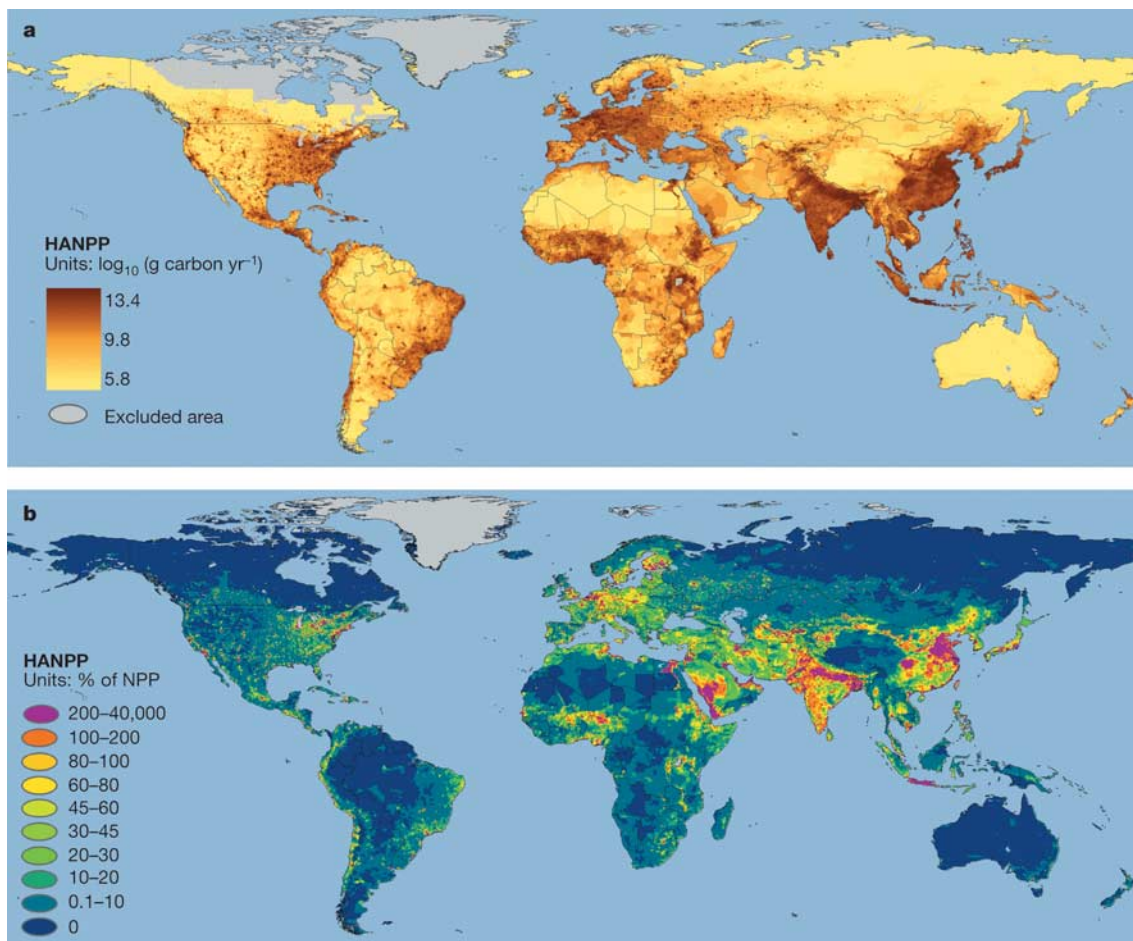


Figure 1 Spatial distribution of the annual NPP resources required by the human population. As measured by **a**, HANPP and **b**, HANPP as a percentage of local NPP. Both maps use the intermediate estimate for HANPP and are in units of carbon.

Table 2 Annual global HANPP estimates

Product	Low estimate (Pg C)	Intermediate estimate (Pg C)	High estimate (Pg C)
Vegetal food	0.90	1.73	2.95
Meat	1.69	1.92	2.22
Milk	0.15	0.27	0.43
Eggs	0.09	0.17	0.26
Food (subtotal)	2.83	4.09	5.86
Paper	0.20	0.28	0.38
Fibre	0.32	0.36	0.42
Wood (fuel)	2.68	4.31	4.71
Wood (construction)	1.97	2.50	3.44
Wood and fibre (subtotal)	5.17	7.45	8.95
Total	8.00	11.54	14.81
% of annual NPP (56.8 Pg)	14.10	20.32	26.07

be 11.5 Pg carbon (24.2 Pg organic matter; 1Pg = 10¹⁵ g) (Table 2). This value is lower than the intermediate estimates of earlier studies^{3,4} (40.6 and 39 Pg organic matter, respectively), but the difference is largely due to items we have omitted. First, we included only the NPP required to produce consumed goods, not the components of NPP that are lost to land transformation (for example, 'shifting cultivation' and 'land clearing'³). Second, we do not treat below-ground biomass in grazed lands as being appropriated by humans because we assume that most of it survives grazing. If we include these components in our analysis, our global estimate increases to 20.8 Pg carbon (43.8 Pg organic matter), a value remarkably close to those reported in the previous studies^{3,4}, despite differences in method.

To address uncertainties we bracketed our estimate of HANPP with low and high calculations using the range of efficiency, loss and residue multipliers reported in the literature (Table 1). For the intermediate estimate we applied these multipliers according to each country's development status¹¹. For low and high estimates we applied to all countries the multipliers that produce minimum and maximum values. Our low and high estimates for HANPP are 8.0 and 14.8 Pg carbon, respectively (Table 2).

To compare HANPP to the global pattern of NPP, we mapped terrestrial NPP using a carbon model driven by global satellite-derived vegetation index and climate data obtained between 1982 and 1998 (refs 13, 14). The use of data collected over such a long baseline reduces short-term variations in NPP while incorporating the cumulative and more recent effects of human influence on the land surface. The resulting global map of NPP matches both the spatial resolution and the time interval of our HANPP data. We estimate global annual NPP to be 56.8 Pg carbon (119.6 Pg organic matter); a value that is close to the average of previous estimates¹⁵.

Expressing HANPP as a percentage of NPP reveals a spatially explicit balance sheet of NPP 'supply' and 'demand' (Fig. 1b), revealing a level of spatial heterogeneity not seen in earlier studies. Globally, we found that humans appropriate approximately 20% of terrestrial NPP, with low and high estimates of 14 and 26%, respectively (Table 2). Some regions, however, such as western Europe and south central Asia, consume more than 70% of their regional NPP (Table 3). Conversely, HANPP in other regions is less

than 15% of NPP, with the lowest value (about 6%) found in South America. At local scales the spatial variation in HANPP is even more striking, varying from nearly 0% of local NPP in sparsely populated areas to over 30,000% in large urban centres (Fig. 1b).

Human populations clearly are not limited to consuming the products of local photosynthesis. Regional and global trade transports these products widely, such that the environmental impacts of human consumption are partly realized far from where products are actually consumed. International trade may also affect HANPP because imported goods are often produced using different efficiencies than those in the consuming country (Table 1). Our high and low estimates set upper and lower boundaries for this latter uncertainty, but our analyses do not explicitly consider transportation issues. Nevertheless, the maps (Fig. 1) identify areas that clearly require net imports of NPP, and they depict the distribution of local impacts accompanying population and consumption (such as waste production, pollution and urbanization). Measuring the global flows of NPP-based goods is a significant future challenge to understanding the environmental impacts of human populations^{16,17}.

The equation $I = PAT$ (ref. 18) describes the overall ecological impact (I) of human activities, which involves the product of population size (P), per capita consumption (A , for 'affluence') and the technologies employed (T). Our results illustrate the importance of each of these factors. The role of population is clear from Fig. 1, despite vast differences in consumption among nations. For example, east and south central Asia, with almost a half of the world's population (Table 3), appropriates 72% of its regional NPP despite having the lowest per capita consumption of any region (1.29 tonnes (t); 1t = 10⁶ g) yr⁻¹). Affluence also plays an important part. Average annual per capita HANPP for industrialized countries (3.2 t) is almost double that of developing nations (1.8 t), which host 83% of the global population¹⁹. If the per capita HANPP of developing nations increased to match that of industrialized countries, global HANPP would increase by 75% to 20.2 Pg carbon (that is, 35% of the current global NPP). Finally, technologies strongly determine the efficiency with which NPP is appropriated. In industrialized countries, for example, 1 t of milled lumber requires approximately 1.3 t of harvestable above-ground tree biomass, whereas in developing countries over 2 t is required (Table 1).

Changing patterns of HANPP will have important consequences for human welfare and global biodiversity. Further growth and development in areas of high HANPP is likely to impoverish local ecosystems and diminish the vital services they provide⁸. These areas will also require increased NPP imports, alterations to flows of NPP-based products and exertions of greater pressure on ecosystems elsewhere^{16,17}. Improved technologies, meanwhile, may help to reduce HANPP through better efficiencies and product substitutions. These changes will be complex, interacting and difficult to predict. Spatially explicit measures of HANPP, however, will help to illuminate current human impacts on the biosphere, monitor changes in these impacts over time and explore the potential of various policies for alleviating them. □

Table 3 HANPP for selected regions of the world (intermediate estimate)

Region*	Area (10 ⁶ km ²)	Population (10 ⁸)	Per capita HANPP (t)	NPP (Pg)	HANPP (Pg)	HANPP (% of NPP)
Africa	31.1	742	2.08	12.50	1.55	12.40
East Asia	11.9	1,400	1.37	3.02	1.91	63.25
South central Asia	10.9	1,360	1.21	2.04	1.64	80.39
Western Europe	1.20	181	2.86	0.72	0.52	72.22
North America	19.7	293	5.40	6.67	1.58	23.69
South America	18.4	316	3.11	16.10	0.98	6.09

* Definitions of regions were taken from ref. 22.

Methods

FAO data

We scanned the country-level FAO data for 1995 for internal consistency, missing data and reporting errors. Missing data were assigned values using the average per capita consumption of countries in the same development category. Over-reporting because of multiple entries for the same country was corrected. For national entities or territories reporting under another administrative country, their populations were added to that of the reporting country to compute the per capita consumption. We defined consumed products as the domestic supply (that is, production plus imports minus exports) to constrain the country totals to products consumed *in situ*.

HANPP calculations

For vegetal foods and fibre, mass was successively added to account for post-harvest processing, transport losses^{20,21} and crop residue²². Crop residue is the residue to product ratio. For the intermediate estimate we used the weighted mean for major world crops whereas high and low estimates are ± 1 s.d. (Table 1). For wood, fuel wood and paper products, organic matter was added to account for processing²³ and harvest losses²⁴. For paper, recycling was also considered²⁵. If the individual plant is killed (all cases except pasture) we included the biomass of the root system using the values in Table 1 (ref. 13).

Meat consumption was based on wet carcass weight and it combined all meat types. The meat component of the total HANPP was estimated by summing the NPP required for grain and pasture-based feed, assuming a global average of 62% grain and 38% forage²⁶. We estimated the amount of organic matter used as feed by applying efficiency values for grain (an average of 2.3:1 kg grain/kg carcass for all meat types) and for pasture (21.46:1 for ruminants) using data from previous studies^{27,28}. The total NPP required for grain feed was then calculated in the same way as for vegetal foods, adding residue and loss factors appropriate to each country's development status. Because grazing occurs *in situ*, no loss or residue factors were added to pasturage. Efficiency factors for milk and eggs are for the grain component only. Carbon/organic matter ratios used for the high, intermediate and low estimates span the range (intermediate estimate uses average value) of reported values for various plants²⁹.

NPP model

We used the Carnegie Ames Stanford Approach carbon model to estimate NPP. The model incorporates satellite and climate data to estimate the fixation and release of carbon based on a spatially and temporally resolved prediction of NPP in a steady state. The performance of this model is evaluated in ref. 15.

Received 6 January; accepted 5 May 2004; doi:10.1038/nature02619.

- Vitousek, P. M., Mooney, H. A., Lubchenco, J. & Melillo, J. M. Human domination of Earth's ecosystems. *Science* **277**, 494–499 (1997).
- Wackernagel, M. *et al.* Tracking the ecological overshoot of the human economy. *Proc. Natl Acad. Sci. USA* **99**, 9266–9271 (2002).
- Vitousek, P. M., Ehrlich, P., Ehrlich, A. & Matson, P. M. Human appropriation of the products of photosynthesis. *Bioscience* **36**, 368–373 (1986).
- Rojstaczer, S., Sterling, S. M. & Moore, N. J. Human appropriation of photosynthesis products. *Science* **294**, 2549–2552 (2001).
- Schimel, D. *et al.* Contribution of increasing CO₂ and climate to carbon storage by ecosystems in the United States. *Science* **287**, 2004–2006 (2000).
- Haberl, H. Human appropriation of net primary production as an environmental indicator: Implications for sustainable development. *Ambio* **26**, 143–146 (1997).
- Field, C. B. Global change: Enhanced: Sharing the garden. *Science* **294**, 2490–2491 (2001).
- Daily, G. C. *et al.* Ecosystem services: benefits supplied to human societies by natural ecosystems. *Issues Ecol.* **2**, 1–16 (1997).
- O'Neill, R. V. & Kahn, J. R. Homo economicus as a keystone species. *Bioscience* **50**, 333–337 (2000).
- Haberl, H., Krausmann, F., Erb, K. H. & Schulz, N. B. Human appropriation of net primary production. *Science* **296**, 1968–1969 (2002).
- Food and Agriculture Organization of the United Nations, *FAOSTAT 2001 Database* (UNFAO, Rome, 2001).
- Center for International Earth Science Information Network, *Gridded Population of the World (GPW) version 2* (Columbia Univ.; International Food Policy Research Institute (IFPRI); and World Resources Institute, New York, 2000).
- Potter, C. S. *et al.* Terrestrial ecosystem production: a process model based on global satellite and surface data. *Glob. Biogeochem. Cycles* **7**, 811–841 (1993).
- Slayback, D. A., Pinzon, J. E., Los, S. O. & Tucker, C. J. Northern hemisphere photosynthetic trends: 1982–99. *Glob. Change Biol.* **9**, 1–15 (2003).
- Cramer, W. *et al.* Comparing global models of terrestrial primary productivity (NPP): overview and key results. *Glob. Change Biol.* **5**(Suppl. 1), 1–15 (1999).
- Luck, M. A., Jenerette, G. D., Wu, J. G. & Grimm, N. B. The urban funnel model and spatially heterogeneous ecological footprint. *Ecosystems* **4**, 782–796 (2001).
- van den Bergh, J. & Verbruggen, H. Spatial sustainability, trade, and indicators: an evaluation of the 'ecological footprint'. *Ecol. Econ.* **29**, 61–72 (1999).
- Holdren, J. P. & Ehrlich, P. R. Human population and global environment. *Am. Sci.* **62**, 282–292 (1974).
- United Nations Population Division *World Population Prospects: The 2000 Revision and The World Population in 2003* (United Nations Population Division Department of Economic and Social Affairs, New York, 2003).
- World Resources Institute *Disappearing Food: How Big are Postharvest Losses?* (World Resources Institute Sustainable Development Information Service, Washington DC, 1998); available at (<http://www.wri.org/trends/foodloss.html>).
- De Padua, D. B. in *Accelerating Agricultural Development* (eds Drilon, J. D. & Sanguiguit, G. F.) 135–154 (SERCA College, Laguna, 1978).
- Smil, V. *Biomass Energies: Resources, Links, Constraints* (Plenum, New York, 1983).

- Ince, P. J. *Industrial Wood Productivity in the United States 1900–1998* (Research Note FPL-RN-0272, USDA Forest Service, Madison, 2000).
- Pulkki, R. E. *Literature Synthesis on Logging Impacts in Moist Tropical Forests* (working paper GFSS/WP/06, Forest Products Division, UNFAO, Rome, 1997).
- Skog, K. E., Ince, P. J. & Haynes, R. W. Wood fiber supply and demand in the United States. *Proc. N. Am. For. Comm.* [online] (<http://www.fpl.fs.fed.us/docmnts/pdf2000/skog00a.pdf>), 1998.
- Sere, C. & Steinfeld, H. *World Livestock Production Systems* (Animal Production and Health Paper 127, FAO, Rome, 1996); available at (<http://www.fao.org/WAICENT/FAOINFO/AGRICULT/AGA/LSPA/Paper127/cover1.htm>).
- Council for Agricultural Science and Technology, *Animal Agriculture and Global Food Supply* (Task Force Report R135, Ames, 1999); available at (http://www.cast-science.org/cast/pub/anag_nr.htm).
- Oltjen, J. W., George, M. R. & Drake, D. J. in *Computers in Agricultural Extension Program* (eds Watson, D. G., Zazueta, F. S. & Bottcher, A. B.) 58–63 (American Society of Agricultural Engineers, St Joseph, 1992).
- Lieth, H. in *Primary Productivity of the Biosphere* (eds Lieth, H. & Whittaker, R. H.) 119–130 (Springer, New York, 1975).

Acknowledgements We thank K. Carney, N. Christiansen, G. Daily, R. Defries, J. Hellmann, J. Hicke, G. Orians, N. Sanders, P. Vitousek, P. Matson and P. Ehrlich for comments, C. Tucker for providing satellite data, and G. Asrar, J. Kaye, G. Gutman and D. Wickland of the US National Aeronautics and Space Administration for financial support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.L.I. (Marc.L.Imhoff@nasa.gov).

A proteoglycan mediates inductive interaction during plant vascular development

Hiroyasu Motose^{1*}, Munetaka Sugiyama² & Hiroo Fukuda^{1,3}

¹Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, Japan

²Botanical Gardens, Graduate School of Science, The University of Tokyo, Hakusan 3-7-1, Bunkyo-ku, Tokyo 112-0001, Japan

³Plant Science Center, RIKEN, Suehiro 1-7-22, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

* Present address: Section of Plant Biology, Division of Biological Sciences, University of California, One Shields Avenue, Davis, California 95616, USA

Inductive cell–cell interactions are essential for controlling cell fate determination in both plants and animals¹; however, the chemical basis of inductive signals in plants remains little understood. A proteoglycan-like factor named xylogen mediates local and inductive cell–cell interactions required for xylem differentiation in *Zinnia* cells cultured *in vitro*^{2,3}. Here we describe the purification of xylogen and cloning of its complementary DNA, and present evidence for its role *in planta*. The polypeptide backbone of xylogen is a hybrid-type molecule with properties of both arabinogalactan proteins and nonspecific lipid-transfer proteins. Xylogen predominantly accumulates in the meristem, procambium and xylem. In the xylem, xylogen has a polar localization in the cell walls of differentiating tracheary elements. Double knockouts of *Arabidopsis* lacking both genes that encode xylogen proteins show defects in vascular development: discontinuous veins, improperly interconnected vessel elements and simplified venation. Our results suggest that the polar secretion of xylogen draws neighbouring cells into the pathway of vascular differentiation to direct continuous vascular development, thereby identifying a molecule that mediates an inductive cell–cell interaction involved in plant tissue differentiation.

The plant vascular system, comprised of xylem and phloem, forms a continuous network throughout the plant body to transport water, nutrients and signalling molecules. Vascular cells differentiate

by contact with each other to form a continuous strand of vasculature. This developmental feature of the vascular system has raised the possibility that some inductive signals function to guide continuous vascular differentiation. Such inductive signals have not been defined, however, in part because of difficulties in analysing cell–cell interactions *in planta*. *In vitro* xylogenesis of *Zinnia* (*Zinnia elegans* L.)^{4,5} offers an excellent experimental system for directly analysing cell–cell interactions and for isolating signal molecules. In this culture system, isolated mesophyll cells of *Zinnia* transdifferentiate into tracheary elements (TEs), distinctive cells composing xylem vessels *in planta* that are characterized by the formation of secondary cell wall thickenings and programmed cell death. Our previous investigation with this culture system suggested that premature TEs draw neighbouring cells into the pathway of TE differentiation². This inductive cell–cell interaction was found to be mediated by a locally acting secretory factor, which was named xylogen for its xylogenetic activity. The activity of xylogen in the conditioned medium of *Zinnia* xylogenetic culture was recovered in the fraction of arabinogalactan proteins (AGPs), a group of plant proteoglycans³.

We purified xylogen from the AGP fraction by using concanavalin A affinity chromatography after a boil treatment. Purified xylogen increased the frequency of TE differentiation in *Zinnia* cell culture, whereas chemically deglycosylated xylogen did not (Fig. 1a–c), indicating that the glycosyl chains of xylogen are required for its activity. Notably, xylogen derived from the dicot *Zinnia* could promote TE differentiation from mesophyll cells of the monocot *Asparagus officinalis* L. (Fig. 1d–f), suggesting that xylogen is not species-specific but is common to diverse angiosperms. SDS polyacrylamide gel electrophoresis (SDS–PAGE) of purified xylogen yielded a broad band with a relative molecular mass of 50,000–100,000 (M_r 50K–100K), whereas deglycosylated xylogen migrated as a sharp band with an M_r of 16K (Fig. 1g). No other protein bands were detected by silver staining, indicating the homogeneity of the purified xylogen. Monoclonal antibody JIM13 and β -glucosyl Yariv reagent (β GlcY), which detect AGPs^{6,7}, bound to purified xylogen but not to deglycosylated xylogen (Fig. 1g), consistent with reports that JIM13 and β GlcY recognize the glycosyl chains of AGPs^{6,8}. Addition of β GlcY to the *Zinnia* cell culture potentially suppressed TE differentiation as compared with α -galactosyl Yariv reagent (α GalY), an inactive analogue of β GlcY (Fig. 1h). From these results, we conclude that xylogen is an AGP that is essential for TE differentiation.

Amino-terminal amino acid sequencing of the 16K deglycosylated xylogen gave the sequence AHQTGAOAOAADCSTVILNM (where O is hydroxyproline). On the basis of this information, a cDNA encoding xylogen was generated by rapid amplification of cDNA ends (RACE). The resulting cDNA had an open reading frame (ORF) of 183 amino acids, which contained the amino acid sequence identical to the N-terminal sequence determined for deglycosylated xylogen (Fig. 2a). The gene represented by this cDNA is designated *Z. elegans* xylogen protein 1 (*ZeXYP1*). Analysis *in silico* indicated that *ZeXYP1* is a hybrid-type molecule with properties of both AGPs and nonspecific lipid-transfer proteins (nsLTPs).

Four domains can be recognized in the sequence of *ZeXYP1* (Fig. 2b): an N-terminal signal peptide, an nsLTP domain, a proline–alanine (Pro–Ala)-rich domain, and a predicted glycosyl phosphatidylinositol (GPI)-anchor attachment site followed by a carboxy-terminal transmembrane region. The nsLTP domain contains eight cysteine residues, which are strictly conserved in plant nsLTPs and form four disulfide bridges^{9,10}. *ZeXYP1* shares two characteristics with the polypeptide backbones of AGPs: a GPI-anchor signal that is conserved in a large family of AGPs^{11–13}, and short runs of proline alternating with alanine or threonine¹⁴ (APAPA just before the nsLTP domain and APTPA in the (Pro–Ala)-rich domain). On the basis of AGP structure¹⁴, two proline residues

of APAPA, which were shown to be hydroxylated, are probably O-linked to arabinogalactans. The binding of xylogen to concanavalin A suggests that an N-glycan side chain containing mannose residues is present at the one potential N-glycosylation site.

Two homologous genes of *Arabidopsis thaliana*, At5g64080 and At2g13820, were identified by a BLAST search with *ZeXYP1* as a query and designated *A. thaliana* xylogen protein 1 (*AtXYP1*) and

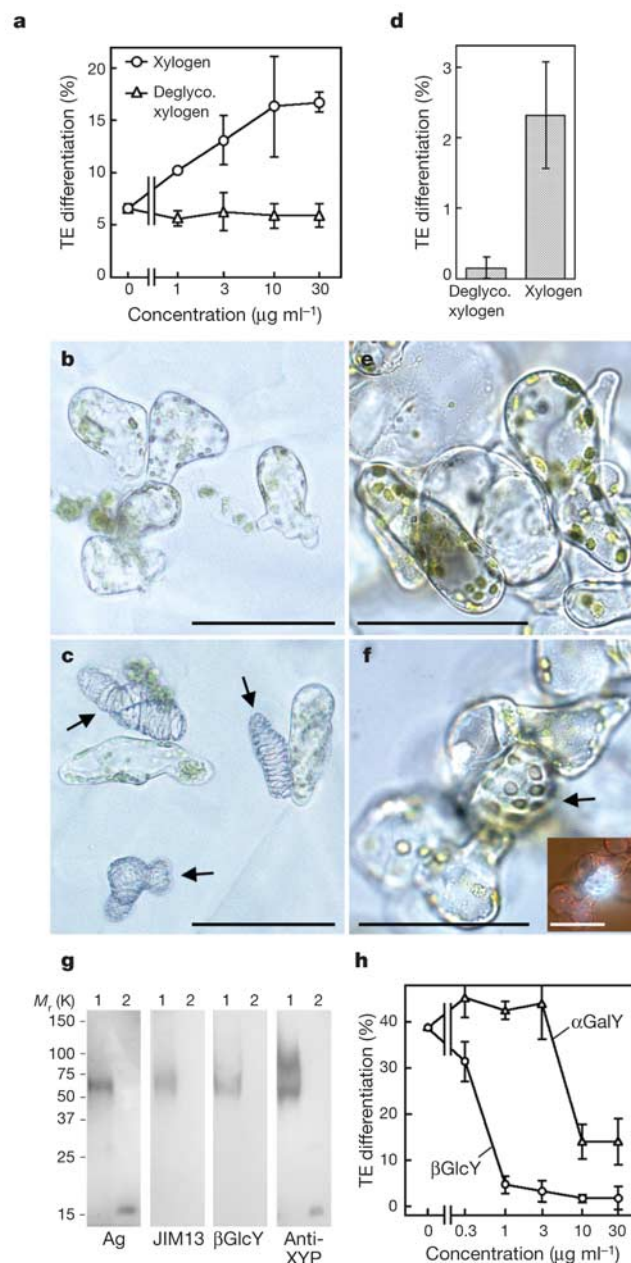
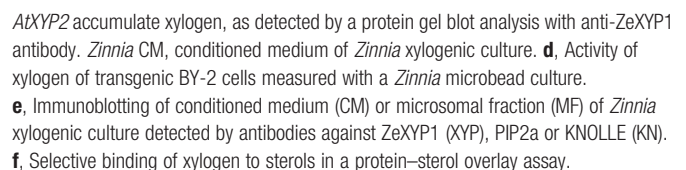


Figure 1 Xylogen is an arabinogalactan protein that induces xylem differentiation. **a–f**, Xylogen induces tracheary element (TE) differentiation of *Zinnia* (**a–c**) and *Asparagus* (**d–f**) mesophyll cells. *Zinnia* cells were cultured with 30 $\mu\text{g ml}^{-1}$ deglycosylated xylogen (**b**) or 30 $\mu\text{g ml}^{-1}$ xylogen (**c**). *Asparagus* cells were cultured with 30 $\mu\text{g ml}^{-1}$ deglycosylated xylogen (**e**) or 30 $\mu\text{g ml}^{-1}$ xylogen (**f**). Arrows indicate TEs (**c, f**). Inset in **f** shows an ultraviolet image indicating lignin deposition in a TE. Scale bar, 50 μm (**b, c, e, f**). **g**, Electrophoretograms of xylogen (lane 1) and deglycosylated xylogen (lane 2). Samples separated by SDS–PAGE were stained with silver (Ag), or blotted onto nitrocellulose membranes and stained with JIM13, β GlcY or anti-*ZeXYP1* antibodies. **h**, β GlcY inhibits TE differentiation in *Zinnia* suspension culture.

We next examined the accumulation pattern of xylogen during TE differentiation of *Zinnia* mesophyll cells (Fig. 3a–d). The *ZeXYP1* transcripts started to accumulate 24 h before visible formation of TEs and peaked at 48 h of culture (Fig. 3a). Immunoblot analysis showed that xylogen accumulated in the culture medium 12 h before visible TE formation (Fig. 3c). The period of accumulation of xylogen corresponds to stage II of xylogenesis⁵, during which dedifferentiated cells restrict their competence and become precursors of TEs. Because auxin and cytokinin are essential for TE differentiation, we evaluated their effects on the accumulation of xylogen. The accumulation of the *ZeXYP1* transcripts was induced



by auxin (Fig. 3a, b), whereas both auxin and cytokinin were required for the accumulation of xylogen (Fig. 3c, d). These results indicate that the accumulation of xylogen is regulated posttranscriptionally by the synergistic action of auxin and cytokinin in association with progression of the TE differentiation program.

Localization of the *ZeXYP1* transcripts and xylogen were investigated by *in situ* hybridization (Fig. 3e–h) and immunohistochemistry (Fig. 3i–m), respectively. The *ZeXYP1* transcripts and xylogen were abundant in procambium and immature xylem cells of the *Zinnia* stems but were hardly detected in interfascicular fibres and phloem cells (Fig. 3e–l). Xylogen started to accumulate in the shoot apical meristem (Fig. 3m). Details of xylogen localization were further analysed by indirect fluorescent antibody technique with longitudinal sections. In the immature xylem, xylogen had a polar localization on the apical side of the cell walls of differentiating TEs (Fig. 3n, o), implying that premature TEs secrete xylogen in a polar manner to act on neighbouring cells directionally. The polar localization of xylogen may involve the function of the GPI anchor, because xylogen proteins have a putative GPI-anchor attachment site and GPI anchors direct an uneven distribution of protein in various GPI-anchored proteins^{16,17}.

Transfer-DNA (T-DNA) insertion mutants of *AtXYP1* or *AtXYP2* were analysed to define the function of xylogen during xylem development *in planta* (Fig. 4). We identified three lines in the

SIGNAL database¹⁸ with T-DNA insertions into *AtXYP1* (salk_073673, salk_103127 and salk_147826) and two lines with insertions into *AtXYP2* (salk_055597 and salk_122768). In addition, *xyp2-3* was identified from the Tag-line pool of the Kazusa DNA Research Institute. Although the locations of the T-DNA inserts were confirmed by polymerase chain reaction (PCR) and sequencing, none of the insertion mutants showed obvious defects in morphology. To determine whether *AtXYP1* and *AtXYP2* function redundantly, we analysed the double knockout mutants *xyp1-1 xyp2-2* and *xyp1-2 xyp2-1*. These mutants showed no detectable accumulation of *AtXYP1* and *AtXYP2* transcripts (Fig. 4b), indicating that the insertions created null mutations. More notably, the double knockouts had obvious morphological defects in vascular development as compared with the wild type (Fig. 4c–n).

The digenic knockouts induced discontinuous and thicker veins (Fig. 4d–f) and the improper interconnection of TEs (Fig. 4h). Some TEs differentiated separately from the xylem vessel networks (Fig. 4h). In addition, there was a reduction in the vein loop architecture in the double mutants. In rosette leaves, the number of tertiary veins was markedly reduced and loop formation was severely impaired, resulting in a simpler venation pattern similar to the open venation of ferns and gymnosperms (Fig. 4i–l). In some seedlings of the double knockouts, disconnection of vessels was found in roots as well as in cotyledons and leaves (Fig. 4m, n).

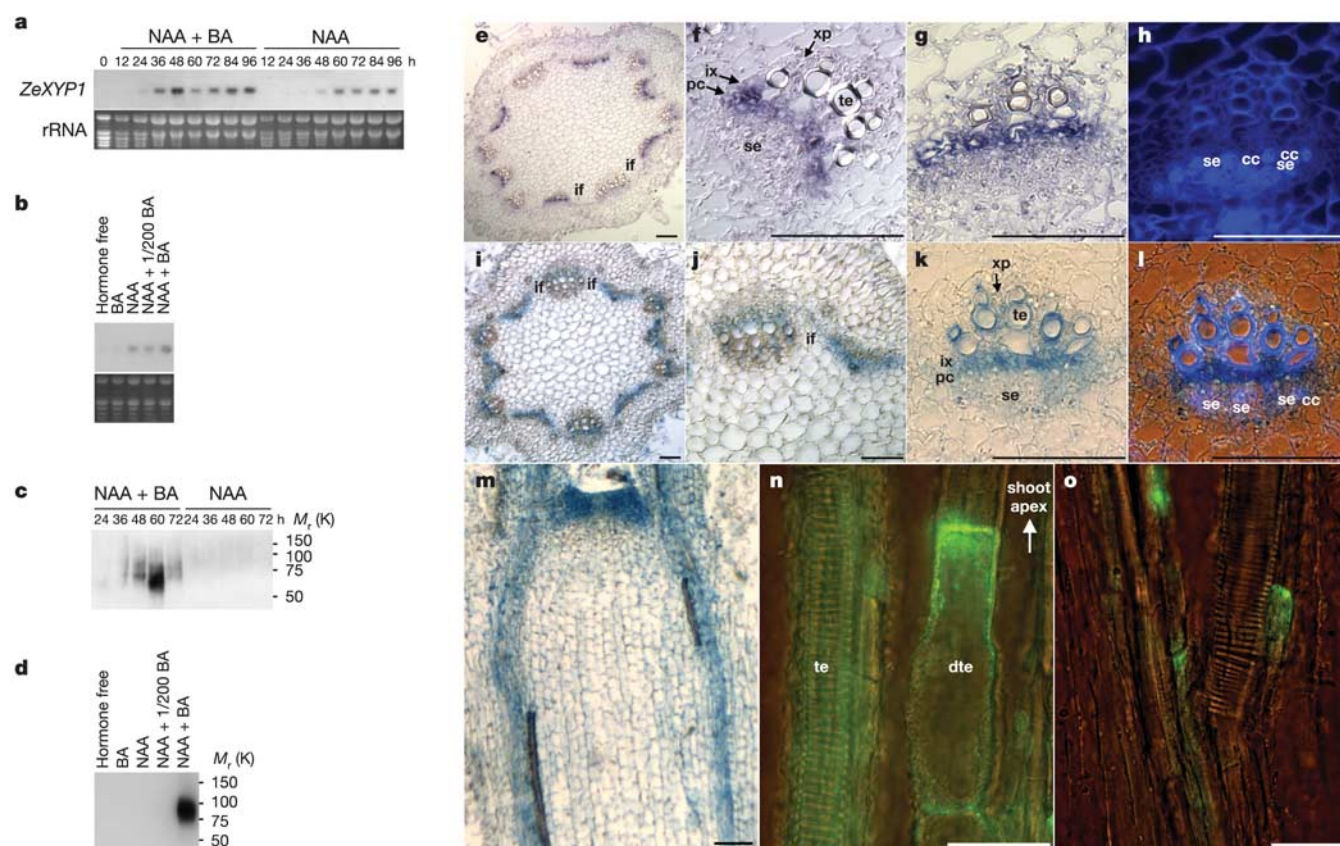


Figure 3 Temporal and spatial profiles of xylogen expression. **a–d**, RNA gel blot analyses of *ZeXYP1* (**a, b**) and protein gel blot analyses of xylogen (**c, d**) in xylogenic (NAA plus BA) and non-xylogenic (hormone free, BA, NAA, NAA plus 1/200 BA) culture. NAA, 0.1 $\mu\text{g ml}^{-1}$ 1-naphthaleneacetic acid; BA, 0.2 $\mu\text{g ml}^{-1}$ 6-benzyladenine; 1/200 BA, 0.001 $\mu\text{g ml}^{-1}$ 6-benzyladenine. **e–h**, *ZeXYP1* transcripts accumulate in procambium and immature xylem cells, as visualized by *in situ* hybridization on a cross-section of a 14-day-old *Zinnia* seedling. **f, g**, Higher magnification images of **e, h**. Ultraviolet image of

g, i–m, Immunohistochemical localization of xylogen in procambium and immature xylem cells of 14-day-old *Zinnia* seedling. **j, k**, Higher magnification images of **i, l**. Ultraviolet image of **k, m**. A longitudinal section. **n, o**, Localization of xylogen by an indirect fluorescent antibody technique. Scale bars, 100 μm (**e–l**), 50 μm (**m**) and 20 μm (**n, o**). cc, companion cell; dte, differentiating TE; if, interfascicular fibre; ix, immature xylem cells; pc, procambium; se, sieve element; te, TE; xp, xylem parenchyma cell.

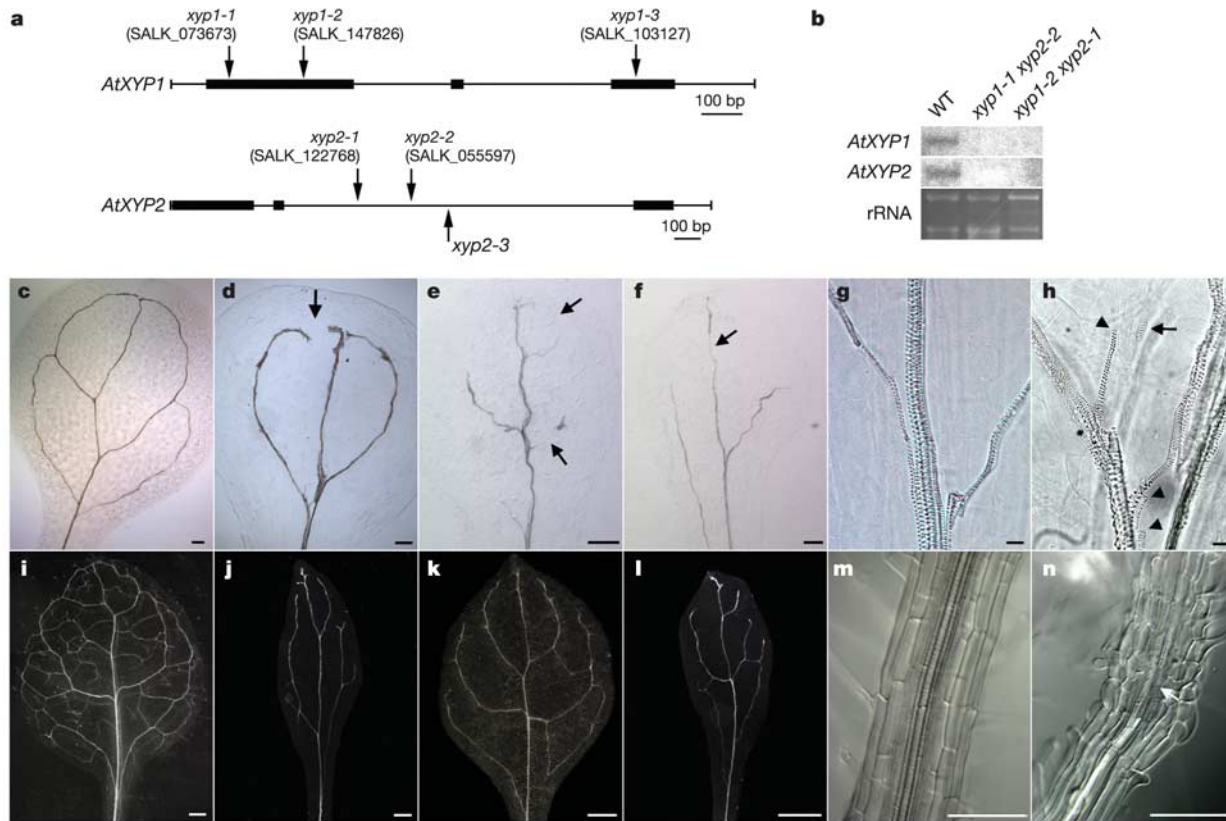


Figure 4 Phenotypic analysis of *Arabidopsis* xylogen knockouts. **a**, Exon–intron structures of *AtXYP1* and *AtXYP2* showing positions of T-DNA in insertion mutants. **b**, The double knockouts (*xyp1-1 xyp2-2* and *xyp1-2 xyp2-1*) completely abolished expression of both *AtXYP1* and *AtXYP2*. WT, Wild type. **c–h**, Vascular patterns of a cotyledon of wild type (**c**, **g**), *xyp1-1 xyp2-2* (**d**, **h**), *xyp1-2 xyp2-1* (**e**) and *xyp1-1 xyp2-3* (**f**) grown for 8 d. **g**, **h**, Higher magnification images of **c**, **d**, respectively. Shown are discontinuous veins (**d–f**, arrows), separated TE (**h**, arrow) and wrongly interconnected vessels (**h**, arrowheads). **i–l**, Vascular patterns of a rosette leaf of wild type (**i**), *xyp1-1 xyp2-2* (**j**), *xyp1-2 xyp2-1* (**k**) and *xyp1-1 xyp2-3* (**l**) grown for 22 d. **m**, **n**, Vascular patterns of a root of wild type (**m**) and *xyp1-1 xyp2-2* (**n**) grown for 8 d. Discontinuous vessel (**n**, arrow). Bars, 200 μ m (**c–f**), 20 μ m (**g**, **h**), 500 μ m (**i–l**) and 100 μ m (**m**, **n**).

Similar phenotypic characters were also observed in *xyp1-1 xyp2-3* (Fig. 4f, l). These results, together with the data obtained with the *in vitro* system, indicate that xylogen may function as a mediator of inductive cell–cell interaction in vascular development.

It should be noted, however, that xylogen seems not to be an essential component of the basic machinery responsible for vascular tissue differentiation, because a deficiency of xylogen did not cause a total loss of vascular tissues but had an influence on limited aspects of vascular development, and in some organs its effect was insignificant. For example, no apparent changes in the vascular structure were detected in hypocotyls of seedlings of the xylogen double knockouts and, in the inflorescence stems of mature plants of the double knockouts, the xylem area was only slightly diminished (data not shown). Thus, we consider that xylogen has the role of a coordinator, rather than a determinant, during vascular development to generate minor veins and to guarantee the integrity of vascular networks.

In summary, our findings on xylogen suggest a scheme in which xylogen is secreted directionally from differentiating vascular cells, moves in the apoplast to the adjacent undifferentiated cells and draws them into the pathway of vascular differentiation to support the formation of continuous networks of vasculature. In addition, the simplified venation in the xylogen double knockouts and the accumulation of xylogen in the meristem suggest that xylogen is involved in the pattern formation of procambial tissues. The phenotype of the xylogen double knockouts reminds us of several *Arabidopsis* mutants that show discontinuous veins^{19,20}. Among these mutants with discontinuous vasculature, *cotyledon vascular*

pattern1 (*cvp1*) is notable in consideration of the function of xylogen²¹. Because *CVPI* encodes sterol methyltransferase 2, an enzyme involved in the biosynthesis pathway of sitosterol and stigmasterol, sterols or sterol-derived molecules are implicated in vascular development as patterning signals²¹. The sterol-binding ability of xylogen, together with the findings on *CVPI*, raises the possibility that xylogen functions as a complex with a type of sterol or sterol-derived molecule. □

Methods

Cell culture

Asparagus mesophyll cells were isolated and cultured as described²². We counted TEs under a light microscope at the tenth day of culture. The frequency of TE differentiation was calculated as the proportion of TEs to the number of living cells. Microbead cultures and suspension cultures of *Zinnia* mesophyll cells were done as described³. The xylogen culture medium for the induction of TE differentiation contained auxin (0.1 μ g ml⁻¹ 1-naphthaleneacetic acid; NAA) and cytokinin (0.2 μ g ml⁻¹ 6-benzyladenine; BA). The non-xylogen media contained four different combinations of auxin and cytokinin: no hormones; 0.2 μ g ml⁻¹ BA; 0.1 μ g ml⁻¹ NAA; or 0.1 μ g ml⁻¹ NAA and 0.001 μ g ml⁻¹ BA.

We monitored the activity of xylogen with microbead cultures³, in which *Zinnia* mesophyll cells were embedded in bead-shaped agarose gels at a cell density of 2.5×10^4 cells per ml (low density) or 8.0×10^5 cells per ml (high density). Low-density and high-density microbeads were cultured together to give a total cell density of 2.2×10^4 cells per ml in the liquid xylogen medium containing various concentrations of xylogen and deglycosylated xylogen. In the suspension cultures, *Zinnia* mesophyll cells were cultured at a cell density of 8.0×10^4 cells per ml in the liquid media. We counted TEs under a light microscope at 96 h of culture. The frequency of TE differentiation was calculated as the proportion of TEs to the number of living cells. Values and error bars are the means and standard deviations, respectively, of triplicate measurements.

To overexpress xylogen proteins, chimaeric genes composed of the cauliflower mosaic virus 35S promoter and the ORF of ZeXYP1, AtXYP1 or AtXYP2 were introduced into tobacco BY-2 cells by *Agrobacterium*-mediated transformation²³. BY-2 cells transformed

with empty vector were used as controls. We isolated AGP fractions from the culture medium of transgenic BY-2 cells by using the specific interaction of AGPs with β GlcY³ and added them into *Zinnia* microbead cultures to monitor the activity of xylogen.

Analysis of xylogen

Xylogen was purified from the AGP fraction, which was obtained from the conditioned medium of the *Zinnia* xylogenic culture by using the specific interaction of AGPs with β GlcY³. The AGP fraction was boiled for 10 min and centrifuged at 10,000 g for 20 min. The supernatant was applied onto concanavalin-A-sepharose column (Amersham Pharmacia Biotech) equilibrated with 10 mM KH₂PO₄-KOH (pH 7.2) containing 150 mM NaCl. The column was washed with the same buffer, and xylogen was eluted with the same buffer containing 0.2 M methyl- α -D-glucoside. Purified xylogen was deglycosylated by trifluoromethanesulphonic acid²⁴. The resulting 16K polypeptide was sequenced by a G1005A Protein Sequencing System (Hewlett Packard) according to the manufacturer's protocol.

For 3'-RACE, the first strand cDNA was reverse-transcribed from poly(A)⁺ RNA prepared from cultured *Zinnia* cells with dT17-LL-BamA primer (GATTAGGATCCAC TAATATCT₁₇). A degenerate primer F1 (GCNCAYCARACIGGIGCICGICCI) was designed for the N-terminal amino acid sequence of the 16K protein backbone of xylogen. A cDNA fragment was amplified from the first strand cDNA by PCR with the primer set F1 and LL-BamA (GATTAGGATCCACTAATATC). DNA of 0.5 kilobase pairs (kbp) was cloned and sequenced. On the basis of the nucleotide sequences of cDNA fragments obtained through 3'-RACE, a backward primer R2 (CCACAATTAGTAGCCAGAAGG AGC) was designed for 5'-RACE. First strand cDNA was synthesized from poly(A)⁺ RNA with the R2 primer, and subjected to dC-tailing reaction. PCR amplification from dC-tailed cDNA was done with a primer set, R2 and poly(G)-AP (GGCCACGCGTCGACT AGTACGGGIIIGGGIIGG) to yield DNA of 0.4 kbp, which was cloned and sequenced.

We carried out the protein-lipid overlay assay as described¹⁵. Nitrocellulose membranes with spots of lipids and sterols were incubated with 0.5 μ g ml⁻¹ xylogen or deglycosylated xylogen. The membranes were incubated first with antibody against ZeXYP1 and then with a peroxide-conjugated secondary antibody.

Expression and localization analysis

RNA gel blot analysis, immunoblotting, *in situ* hybridization and immunohistochemistry were done as described²⁵. A rabbit antibody against ZeXYP1 was generated against a synthetic peptide of ZeXYP1 (CLSYVTAGSTVKKPEGT) and purified with affinity chromatography. Monoclonal antibody JIM13 was a gift from K. Roberts (John Innes Centre). We purchased polyclonal antibodies against plasma membrane intrinsic protein 2a (PIP2a) and KNOLLE from Rose Biotechnology. Alkaline phosphatase (AP)-labelled dextran polymer conjugated with anti-rabbit Ig (EnVision/AP, DakoCytomation) and fluorescein isothiocyanate (FITC)-conjugated anti-rabbit Ig (Sigma-Aldrich) were used as the secondary antibodies. The antigen-EnVision/AP complex was detected by an AP Substrate kit (Vector Laboratories).

T-DNA insertion mutants

xyp1-1 (salk_073673), *xyp1-2* (salk_147826), *xyp1-3* (salk_103127), *xyp2-1* (salk_122768) and *xyp2-2* (salk_055597) were provided from SIGnAL. Information about these mutants was obtained from the SIGnAL website (<http://signal.salk.edu>). *xyp2-3* (the Landbeg *erecta* ecotype), identified from the Tag-line pool of the Kazusa DNA Research Institute, was a gift from A. Nagatani (Kyoto University). We analysed the vascular patterns of wild type and T-DNA mutants as described²⁰.

Received 19 January; accepted 4 May 2004; doi:10.1038/nature02613.

- Van Den Berg, C., Willemsen, V., Hage, W., Weisbeek, P. & Scheres, B. Cell fate in the *Arabidopsis* root meristem determined by directional signalling. *Nature* **378**, 62–65 (1995).
- Motose, H., Fukuda, H. & Sugiyama, M. Involvement of local intercellular communication in the differentiation of zinnia mesophyll cells into tracheary elements. *Planta* **213**, 121–131 (2001).
- Motose, H., Sugiyama, M. & Fukuda, H. An arabinogalactan protein(s) is a key component of a fraction that mediates local intercellular communication involved in tracheary element differentiation of zinnia mesophyll cells. *Plant Cell Physiol.* **42**, 129–137 (2001).
- Fukuda, H. & Komamine, A. Establishment of an experimental system for the study of tracheary element differentiation from a single cell isolated from the mesophyll of *Zinnia elegans*. *Plant Physiol.* **65**, 57–60 (1980).
- Fukuda, H. Tracheary element differentiation. *Plant Cell* **9**, 1147–1156 (1997).
- Yates, E. A. *et al.* Characterization of carbohydrate structural features recognized by anti-arabinogalactan-protein monoclonal antibodies. *Glycobiology* **6**, 131–139 (1996).
- Anderson, R. L., Clarke, A. E., Jermyn, M. A., Knox, R. B. & Stone, B. A. A carbohydrate-binding arabinogalactan protein from liquid suspension cultures of endosperm from *Lolium multiflorum*. *Aust. J. Plant Physiol.* **4**, 143–158 (1977).
- Komalavilas, P., Zhu, J.-K. & Nothnagel, E. A. Arabinogalactan-proteins from the suspension culture medium and plasma membrane of rose cells. *J. Biol. Chem.* **266**, 15956–15965 (1991).
- Yamada, M. Lipid transfer proteins in plants and microorganisms. *Plant Cell Physiol.* **33**, 1–6 (1992).
- Kader, J. C. Lipid-transfer proteins in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47**, 627–654 (1996).
- Youl, J. J., Bacic, A. & Oxley, D. Arabinogalactan-proteins from *Nicotiana glauca* and *Pyrus communis* contain glycosylphosphatidylinositol membrane anchors. *Proc. Natl Acad. Sci. USA* **95**, 7921–7926 (1998).
- Oxley, D. & Bacic, A. Structure of the glycosylphosphatidylinositol anchor of an arabinogalactan protein from *Pyrus communis* suspension-cultured cells. *Proc. Natl Acad. Sci. USA* **96**, 14246–14251 (1999).
- Svetek, J., Yadav, M. P. & Nothnagel, E. A. Presence of a glycosylphosphatidylinositol lipid anchor on rose arabinogalactan proteins. *J. Biol. Chem.* **274**, 14724–14733 (1999).

- Nothnagel, E. A. Proteoglycans and related components in plant cells. *Int. Rev. Cytol.* **174**, 195–291 (1997).
- Stevenson, J. M., Perera, I. Y. & Boss, W. F. A phosphatidylinositol 4-kinase pleckstrin homology domain that binds phosphatidylinositol 4-monophosphate. *J. Biol. Chem.* **273**, 22761–22767 (1998).
- Rodriguez-Boulant, E. & Powell, S. K. Polarity of epithelial and neuronal cells. *Annu. Rev. Cell Biol.* **8**, 395–427 (1992).
- Schindelman, G. *et al.* COBRA encodes a putative GPI-anchored protein, which is polarly localized and necessary for oriented cell expansion in *Arabidopsis*. *Genes Dev.* **15**, 1115–1127 (2001).
- Alonso, J. M. *et al.* Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301**, 653–657 (2003).
- Berleth, T. & Mattsson, J. Vascular development: tracing signals along veins. *Curr. Opin. Plant Biol.* **3**, 406–411 (2000).
- Koizumi, K., Sugiyama, M. & Fukuda, H. A series of novel mutants of *Arabidopsis thaliana* that are defective in the formation of continuous vascular network: calling the auxin signal flow canalization hypothesis into question. *Development* **127**, 3197–3204 (2000).
- Carland, F. M., Fujioka, S., Takatsuto, S., Yoshida, S. & Nelson, T. The identification of CVP1 reveals a role for sterols in vascular patterning. *Plant Cell* **14**, 2045–2058 (2002).
- Matsubayashi, Y. & Sakagami, Y. Phytosulfokine, sulfated peptides that induce the proliferation of single mesophyll cells of *Asparagus officinalis* L. *Proc. Natl Acad. Sci. USA* **93**, 7623–7627 (1996).
- Ito, M. *et al.* A novel *cis*-acting element in promoters of plant B-type cyclin genes activates M phase-specific transcription. *Plant Cell* **10**, 331–341 (1998).
- Sojar, H. T. & Bahl, O. P. Chemical deglycosylation of glycoproteins. *Methods Enzymol.* **138**, 341–359 (1987).
- Nishitani, C., Demura, T. & Fukuda, H. Primary phloem-specific expression of a *Zinnia elegans* Homeobox gene. *Plant Cell Physiol.* **42**, 1210–1218 (2001).

Acknowledgements We thank J. J. Harada for critically reading the manuscript; K. Roberts for JIM13; S. Hasegawa and T. Nagata for tobacco BY-2 cells; S. Sawa for advice on the protein-lipid overlay assay; N. Sassa, K. Iwamoto and C. Nishitani for advice on *in situ* hybridization; M. Fujita and N. Shinohara for advice on immunohistochemistry; the Salk Institute Genomic Analysis Laboratory for the sequence-indexed *Arabidopsis* T-DNA insertion mutants; the Kazusa DNA Research Institute for the Tag-line pool; A. Nagatani for *xyp2-3* seeds; K. Ohashi-Ito and S. Sakamoto for the screening and isolation of *xyp2-3*. H.M. was supported in part by the Yamada Science Foundation. Funding for the SIGnAL indexed insertion mutant collection was provided by the National Science Foundation of USA. This work was supported in part by Grants-in-Aid from the Ministry of Education, Science, Sports and Culture of Japan, from the Japan Society for the Promotion of Science, and from the Mitsubishi Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.M. (hmotose@ucdavis.edu). The sequence for ZeXYP1 is deposited in GenBank under accession number AB159560.

Rapid BDNF-induced retrograde synaptic modification in a developing retinotectal system

Jiu-lin Du & Mu-ming Poo

Division of Neurobiology, Department of Molecular and Cell Biology, Helen Wills Neuroscience Institute, University of California, Berkeley, California 94720, USA

In cultures of hippocampal neurons, induction of long-term synaptic potentiation or depression by repetitive synaptic activity is accompanied by a retrograde spread of potentiation or depression, respectively, from the site of induction at the axonal outputs to the input synapses on the dendrites of the presynaptic neuron^{1,2}. We report here that rapid retrograde synaptic modification also exists in an intact developing retinotectal system. Local application of brain-derived neurotrophic factor (BDNF) to the *Xenopus laevis* optic tectum, which induced persistent potentiation of retinotectal synapses, led to a rapid modification of synaptic inputs at the dendrites of retinal ganglion cells (RGCs), as shown by a persistent enhancement of light-evoked excitatory synaptic currents and spiking activity of RGCs. This retrograde effect required TrkB receptor activation, phospholipase C γ activity and Ca²⁺ elevation in RGCs, and was accounted for by a selective increase in the number of postsynaptic AMPA-subtype glutamate receptors at RGC

dendrites. Such retrograde information flow in the neuron allows rapid regulation of synaptic inputs at the dendrite in accordance to signals received at axon terminals, a process reminiscent of back-propagation algorithm for learning in neural networks³.

To investigate whether retrograde synaptic modulation occurs *in vivo*, we examined whether potentiation of retinotectal connections induced by BDNF, a neurotrophin expressed in the tectum⁴, is accompanied by changes of synaptic inputs at RGC dendrites. Using loose patch stimulation of RGCs and whole-cell perforated patch recording from the tectal neuron of developing *Xenopus* tadpoles (stage 42–45; see Methods), we found that the amplitude of evoked monosynaptic excitatory postsynaptic currents (EPSCs) increased within 5 min of local BDNF application in the tectum (Fig. 1a). Analysis of miniature EPSCs (mEPSCs) in the presence of tetrodotoxin (TTX) showed that the frequency of mEPSCs was also elevated in all cells examined (Fig. 1b, c), whereas the mean mEPSC amplitude remained unchanged (Fig. 1d). The BDNF effects were abolished by the pre-treatment (30 min) of the tectum with k252a, a tyrosine kinase inhibitor⁵, suggesting that receptor tyrosine kinase TrkB is involved (Fig. 1a, c). Thus, the potentiation of retinotectal synapses by BDNF seemed to result from a TrkB-mediated enhancement of presynaptic transmitter release rather than an increase in the postsynaptic responses, consistent with previous findings in other systems^{6–10}.

To assay synaptic inputs at RGC dendrites, we measured light-evoked responses in RGCs using whole-cell perforated patch recording. Briefly, whole-field illumination of the retina elicited both ON and OFF responses in most RGCs (116 out of 132), ON-only responses in 10 out of 132 cells and OFF-only responses in 6 out of 132 cells. Compound synaptic currents (CSCs) evoked by the light stimulus were monitored at two different clamping voltages, corresponding to the reversal potentials (E_e and E_i) of glutamatergic and γ -aminobutyric acid (GABA)/glycinergic synaptic currents, respectively (Fig. 2a). Excitatory CSCs evoked by either the onset or offset of the stimulus were followed by inhibitory CSCs, with a delay of 30 to 200 ms. The strength of excitatory inputs to the RGC is estimated by the total integrated charge associated with CSCs within a 100 ms window after the onset of excitatory inputs, when RGCs were held at E_i (Fig. 2a). Light-evoked excitatory CSCs remained stable when tested at a 60-s interval (Fig. 2b). However, we observed a marked enhancement of excitatory CSCs for both ON and OFF responses within 20 min after intravitreal BDNF application

near the tectum (Fig. 2c, d). A significant increase in the amplitude of spontaneous EPSCs (sEPSCs) was also observed after the light responses (see sample traces in Fig. 2c, d). The effect of potentiation persisted for as long as stable recording could be made (up to 3 h). Because most processes of RGCs and tectal cells are below the surface of the tectum, the actual BDNF concentration may be lower because of dilution and tissue penetration. Experimentally, no significant effects were detected when the total amount of BDNF applied was less than 40 ng (see Supplementary Fig. 1). In this work, a dose of 400 ng BDNF was used in all experiments. As similar BDNF effects on excitatory inputs were found in all RGC subtypes in these tadpoles, the data were pooled.

The enhancement of RGC light responses was not owing to the spread of BDNF from the tectum to the retina. In tadpoles for which the optic nerve of the recorded retina was severed or the tectum was surgically removed, we found that BDNF application had no effect on RGC light responses (Fig. 2c, d). Furthermore, BDNF application at the retina had no effect ($n = 7$, data not shown), suggesting that the BDNF action is restricted to the RGC axon terminals in the tectum. Tectal application of nerve growth factor (NGF) and neurotrophin-3 (NT-3) (two other neurotrophins⁵) at the same concentration as BDNF, had no detectable effect on the RGC light responses (see Supplementary Fig. 2), further indicating the specificity of the BDNF effect.

The enhanced light responses may be caused by changes in the efficacy of bipolar cell synapses on the RGC dendrites. The mean sEPSC amplitude observed in RGCs in the absence of light stimuli exhibited a marked increase after tectal application of BDNF, with a time course similar to that of the changes in light-evoked excitatory CSCs, whereas no significant increase in the sEPSC frequency was observed (Fig. 3a). This increase in the mean sEPSC amplitude was not observed when the optic nerve was severed (Fig. 3a). These sEPSCs were mediated by the AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) subtype of glutamate receptors, because they were abolished by AMPA receptor (AMPA) antagonist CNQX (20 μ M) ($n = 5$, data not shown). There was no change in the 10–90% rise time of sEPSCs (0.53 ± 0.06 ms versus 0.52 ± 0.04 ms) and decay time (1.33 ± 0.13 ms versus 1.35 ± 0.11 ms) ($n = 6$), suggesting no significant change in the properties of the AMPAR channel.

Analysis of the current noise associated with the decay phase of sEPSCs further indicated that the increase in sEPSC amplitude was

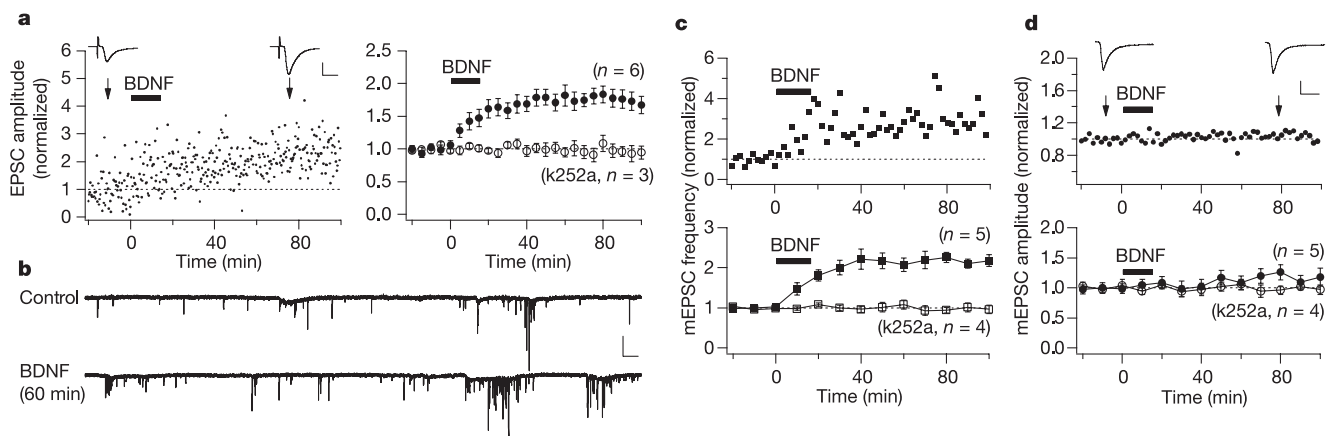


Figure 1 Potentiation of retinotectal synapses by BDNF application at the tectum. **a**, Changes in the amplitude of EPSCs after tectal application of BDNF (bar). An example (left, $V_c = -65$ mV, in a stage 44 tadpole) and summary (right). The data were normalized by the mean value (dotted line) observed before the BDNF application. Sample EPSCs (left) are averages of 15 events (arrow). Scale bars: 40 pA (vertical); 10 ms (horizontal). **b**, Examples of recordings ($V_c = -65$ mV) of spontaneous events (in the

presence of TTX) in a tectal neuron before and 60 min after tectal application of BDNF. Scale bars: 20 pA; 300 ms. **c**, **d**, Changes in the mean mEPSC frequency (**c**) and amplitude (**d**) after tectal application of BDNF (bar). An example (top) and summary (bottom) are shown. The data was normalized to that observed before the BDNF application. Sample mEPSCs are shown in **d**, averaged over a 2 min period (arrow). Scale bars: 3 pA; 10 ms. Pretreatment with k252a (open symbols), untreated (filled symbols).

because of an increase in the number rather than the weighted single-channel conductance or mean channel open probability of synaptic AMPARs (Fig. 3b, c). Similar results were obtained when we analysed mEPSCs in the presence of TTX (Fig. 3c), which showed similar amplitudes and kinetics with those of sEPSCs. The reliability of the noise analysis was confirmed by positive control experiments in which the number of functional AMPARs was reduced by a low concentration of CNQX (0.05–0.15 μ M, see Supplementary Fig. 3). By monitoring sEPSCs in Mg^{2+} -free and glycine-containing external solution, we also examined the

amplitude of *N*-methyl-D-aspartate receptor (NMDAR) component of sEPSCs before and after the tectal BDNF application. After BDNF application, the NMDAR component, which was abolished by *D*(-)-2-amino-5-phosphonovaleric acid (AP5; 50 μ M) ($n = 4$, data not shown), was not changed, whereas the AMPAR component was substantially increased (Fig. 3d). Furthermore, this increase in AMPAR/NMDAR was not observed when the optic nerve was severed (Fig. 3e). Thus the enhancement of sEPSCs was due to an increase in the number of postsynaptic AMPARs at RGC dendrites, similar to that accompanying activity-induced

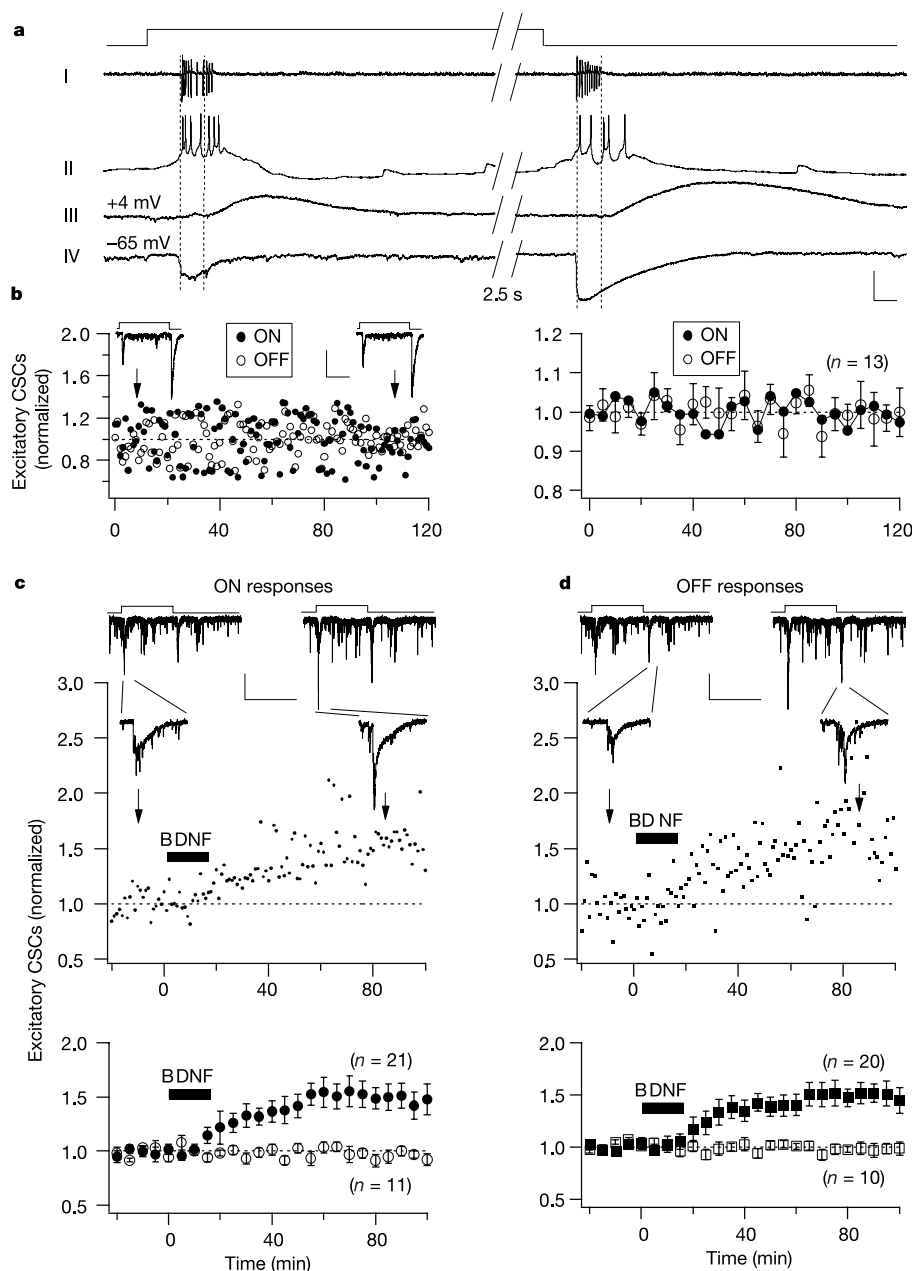


Figure 2 Potentiation of light-evoked excitatory CSCs recorded in RGCs by tectal application of BDNF. **a**, An example of sequential recording from a RGC (in a stage 43 tadpole) in response to light stimulus (4 s, top trace): (trace I) spikes detected under cell-attached mode, (II) membrane potential changes under current clamp mode, and (III and IV) IPSCs and EPSCs recorded under voltage clamp mode at $V_c = +4$ mV and -65 mV, respectively. Scale bars: (vertical) 100 pA (I), 50 mV (II) or 60 pA (III and IV); (horizontal) 100 ms. **b**, Long-term recording showing the stability of RGC light responses. An example (left) and summary (right) of changes in the integrated charge of light-evoked excitatory CSCs, elicited by whole-field light stimuli (4 s, 0.02 Hz). The integrated charge was

normalized by the mean value (dotted line) observed during the first 20 min. Sample traces shown above. Scale bars: 40 pA; 2 s. **c, d**, Potentiation of light-evoked ON (**c**) and OFF (**d**) RGC responses after tectal application of BDNF (bar). An example of data from a RGC ($V_c = -67$ mV) in a stage 44 tadpole (Top). Sample responses shown above. Scale bars: 30 pA; 4 s, 300 ms (for enlarged trace). Summary of results (Bottom) showing BDNF-induced enhancement of ON and OFF responses (filled symbols), and no BDNF effect when the optic nerve was severed or the tectum was surgically removed (open symbols).

long-term potentiation (LTP)^{11,12} or synapse maturation¹³. The characteristic time for BDNF-induced increase in light-evoked RGC responses (20 min) was longer than that of BDNF-induced potentiation of retinotectal synapses (5 min), consistent with the time required for retrograde signalling from RGC axonal terminals to the dendrite. Given the average length of RGC axons, in the range of 0.2–0.5 mm for stage 42–45 tadpoles, the speed of retrograde signal propagation is about 0.2–0.5 $\mu\text{m s}^{-1}$, within the range of estimated rates of vesicular retrograde transport¹⁴.

To further determine the physiological consequence of the retrograde BDNF signalling in RGCs, light-evoked spiking activity of RGCs was monitored by cell-attached patch recording. The spike rate associated with both ON and OFF responses was significantly increased after the tectal BDNF application, with a time course similar to that of the changes in light-evoked excitatory CSCs (see Supplementary Fig. 4a, b). This was due to an increase in the mean firing frequency rather than the duration of spike train evoked by

the light stimulus (see Supplementary Fig. 4c), and was abolished by severing the optic nerve (see Supplementary Fig. 4a, b). Spontaneous spiking activity in the absence of light stimuli also showed a marked increase of the mean firing frequency after the BDNF application (see Supplementary Fig. 5). This increased firing seemed to result largely from more frequent spontaneous bursting and higher number of spikes within each burst, and the effect was abolished by severing the optic nerve. Thus, retrograde synaptic modification of input synapses of RGCs had indeed modulated the spiking output of RGCs.

To further explore the mechanism of the retrograde BDNF effect, we investigated whether TrkB signalling in the RGC is required for the retrograde synaptic modification. A morpholino knockdown approach was used to specifically reduce the expression of TrkB in the retina of one eye and its ipsilateral tectum. This was achieved by the injection of anti-*Xenopus* TrkB (anti-xTrkB) morpholino oligonucleotides into one of the dorsal blastomeres at the animal pole of

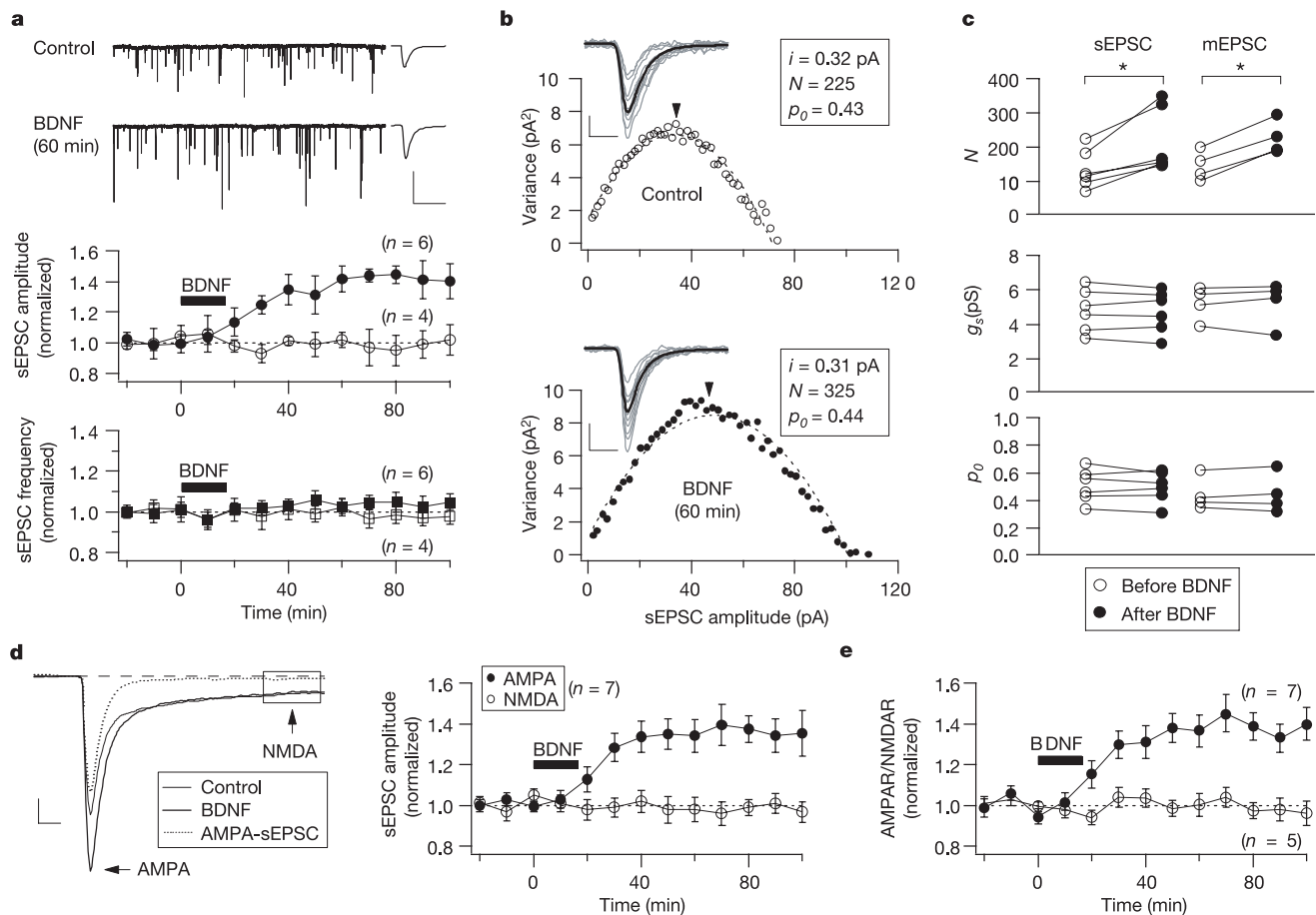


Figure 3 Effects of the tectal application of BDNF on sEPSCs and mEPSCs in RGCs. **a**, Increase of the sEPSC amplitude without any change in sEPSC frequency after tectal application of BDNF (bar). Top, examples of recordings (left) and averaged sEPSCs (right). Scale bars: 50 pA, 1 s (recording) or 6 ms (averaged). Middle and bottom, summary of results showing BDNF-induced changes in the mean sEPSC amplitude and frequency without (filled symbols) and with (open symbols) transection of the optic nerve. **b**, Noise analysis of sEPSCs. An example of analysis for sEPSCs over a 20-min period before (top) and 60 min after BDNF application (bottom) (see Methods). The best fit yields a weighted single-channel current (i), mean channel number (N), and mean channel open probability (p_o), with that corresponding calculated weighted single-channel conductance (g_s) was 4.61 and 4.46 pS, respectively, before and after BDNF application. The arrowhead marks the amplitude of the mean sEPSCs. Sample (grey) and mean (black) sEPSC traces shown above. Scale bars: 10 pA (top) or 20 pA (bottom); 2 ms. **c**, Summary of changes in N , g_s , and p_o associated with sEPSCs ($n = 6$) and mEPSCs ($n = 4$) after the BDNF application.

Values derived from the same RGC are connected by a line (* $P < 0.01$, paired Student's t -test). **d**, Selective increase in the AMPAR component of sEPSCs after the tectal application of BDNF. Average traces of sEPSCs obtained from a RGC ($V_c = -65$ mV) over a 20-min period before (thin) and 60 min after (thick) tectal application of BDNF, recorded in Mg^{2+} -free and glycine (10 μM)-containing solution (left). The peak amplitude and average amplitude at 15–20 ms after the peak were taken as the AMPAR- and NMDAR-mediated component of sEPSCs, respectively. The dotted trace shows the mean sEPSC derived from the control data shown in **a**, which is mediated by AMPARs and rapidly decays to baseline level (dashed line). Scale bars: 4 pA; 2 ms. Summary of changes in the amplitude of AMPAR- and NMDAR-mediated components of sEPSCs after the BDNF application (right), normalized to the control value. **e**, Changes of the ratio of AMPAR to NMDAR component of sEPSCs after the BDNF application (same data set as in **d**). Optic nerve intact (filled symbols), optic nerve severed (open symbols).

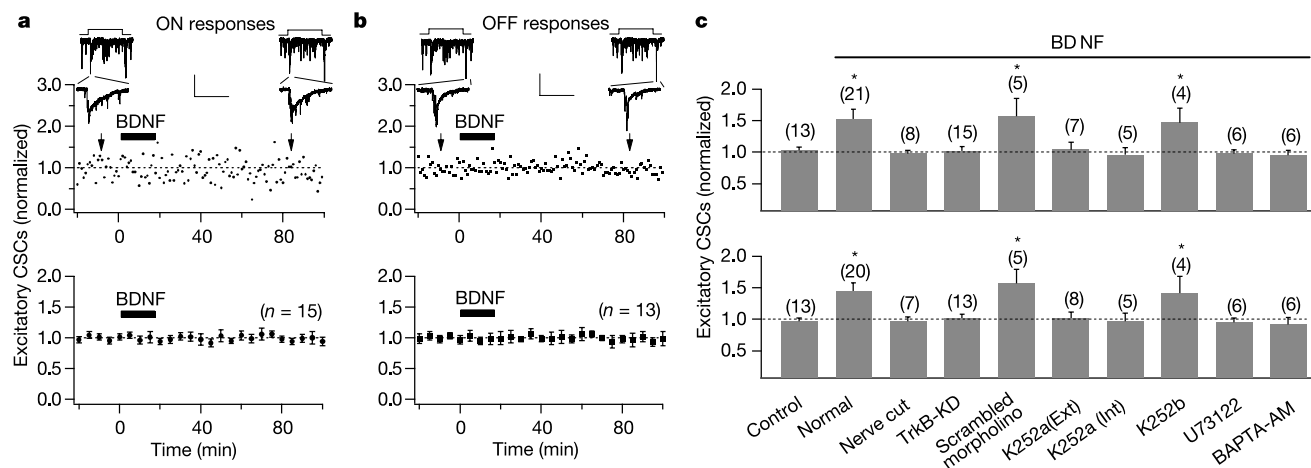


Figure 4 Mechanisms underlying the retrograde synaptic modulation by BDNF. **a, b**, Morpholino knockdown of TrkB in the retina abolished the retrograde effect of BDNF (bar). An example of an experiment (top) and summary of results (bottom). The light-evoked excitatory CSCs were recorded from a TrkB-KD RGC ($V_c = -70$ mV) in a stage 44 tadpole (see Methods). Scale bars: 30 pA; 4 s or 300 ms (for enlarged trace). **c**, Summary of all results. The mean integrated charges of ON (top) and OFF (bottom) responses of light-evoked excitatory CSCs at $t = 80$ – 100 min were normalized by the mean values before the BDNF application. Control, data from figure 2b. Normal and Nerve

Cut, data from figure 2c and d. TrkB-KD, data shown in **a** and **b**. Scrambled morpholino data from experiments similar to that in **a** and **b**, except oligonucleotides of a scrambled sequence were injected into the *Xenopus* embryo. Tadpoles were pre-incubated (30 min) in bath solution containing k252a (labelled k252a(Ext); 200 nM), k252b (200 nM) or U73122 (5 μ M) before CSCs recorded. RGCs were loaded (intracellularly) with k252a (labelled k252a(Int); 200 nM) or BAPTA-AM (2 mM) before CSCs recorded. Number of experiments shown in brackets (* $P < 0.01$, paired Student's *t*-test).

eight-cell stage embryos (see Methods). For the retina with TrkB knockdown (TrkB-KD), BDNF application near the contralateral tectum had no obvious effects on both ON and OFF light-evoked RGC responses (Fig. 4a, b), whereas enhanced RGC responses were observed after BDNF application when control scrambled-sequence morpholino oligonucleotides were used (Fig. 4c). In normal uninjected tadpoles, pre-incubation of the tadpole in the bath solution containing k252a (200 nM) or intracellularly loaded k252a (200 nM; loaded into the RGCs through whole-cell recording pipettes; see Supplementary Fig. 6) abolished the BDNF effect on RGC light responses. In contrast, pre-incubation of the tadpole with k252b (200 nM), a compound that does not inhibit tyrosine kinase at the concentration used, did not change the BDNF effect (Fig. 4c). These results, together with that found in retinas with severed optic nerves, suggest that the retrograde effect of BDNF was mediated by TrkB located at the RGC axon terminals. Furthermore, when two downstream events of TrkB signalling⁵ (phospholipase C γ (PLC- γ) activation and Ca^{2+} elevation) were blocked by pre-treatment with a specific PLC- γ inhibitor U73122 (5 μ M) and intracellular loading of BAPTA-AM (2 mM), respectively, the effect of BDNF in enhancing light-evoked RGC responses was abolished (Fig. 4c).

Target-derived trophic factors, including neurotrophins, are known to be retrogradely transported from axonal terminals to the soma^{5,10,15,16} and are required for the maintenance of neuronal function^{5,10}, including the integrity of synapses at the dendrite^{17,18}. Axotomy of postganglionic axons of sympathetic ganglion neurons led to a severe depression of preganglionic synaptic inputs and withdrawal of synaptic contacts within 3–11 days¹⁷. Interestingly, an exogenous supply of NGF prevents synapse disintegration in sympathetic ganglia¹⁸. In the present study, we found a rapid neurotrophin-induced retrograde synaptic modification that occurs in the order of tens of minutes and involves a selective increase in synaptic AMPARs at RGC dendrites. This leads to an accelerated maturation of glutamatergic synapses on RGC dendrites and enhanced light responses. As the blockade of excitatory synaptic inputs decreases the dendritic motility of RGCs (ref. 19), the BDNF-induced retrograde synaptic modification shown here may be causally linked to structural modification of RGC dendrites observed hours after manipulating the level of tectal BDNF²⁰.

Furthermore, retrograde BDNF signalling increases light-evoked as well as spontaneous spiking activity, resulting in elevated tectal activity. Given that the expression, secretion and trafficking of BDNF are activity-dependent^{10,21–23}, the retrograde BDNF effects on RGCs may lead to elevated BDNF secretion at the tectum, providing a positive-feedback mechanism for activity-dependent maturation of the retinotectal system²⁴. Tectal BDNF messenger RNA peaks at stage 45/46 tadpoles⁴, when active arborization of RGC axons and tectal cell dendrites occurs^{4,20}. The presence of retrograde BDNF signalling in RGCs may ensure coordinated dendritic maturation in the tectum and retina, allowing central regulation of retinal development²⁵. Finally, in functional neural circuits, rapid retrograde synaptic modification also allows coordinated adjustments of synaptic inputs based on signals received at the axonal terminals of a neuron, a mechanism reminiscent of back-propagation algorithms used for supervised learning in artificial neural networks^{3,26}. □

Methods

Tadpole preparation

Xenopus laevis tadpoles were grown at room temperature (22°C), with a 12 h shift in light–dark exposure. For recording, stage 42–45 tadpoles²⁷ were anesthetized with saline containing 0.02% MS222 (Sigma) and secured by insect pins to a sylgard-coated dish and incubated in HEPES-buffered saline (115 mM NaCl, 2 mM KCl, 10 mM HEPES at pH 7.4, 3 mM $CaCl_2$, 10 mM glucose and 1.5 mM $MgCl_2$). Stimulation and recording conditions have been described previously²⁸.

Generation of tadpoles with RGC TrkB knockdown

An antisense technique was used to downregulate the expression level of *Xenopus* TrkB (xTrkB) in the retina and the ipsilateral tectum. Morpholino oligonucleotides downregulate gene expression by inhibiting translation initiation through its binding to the 5' UTR of the target mRNA. In order to knockdown the expression level of xTrkB in the developing *Xenopus* retinotectal system, anti-xTrkB morpholino oligonucleotide (5'-CCACTGGATCCCCCTAGATGGAG-3'), which is specific to xTrkB, and the control morpholino oligonucleotide of a scrambled sequence (5'-CCTCTTACCTCAGTTA CAATTATATA-3') were synthesized by Gene Tools, LLC, and resuspended in sterile water. The morpholino oligonucleotide (2 ng) was injected into one dorsal blastomere in the animal pole of an eight-cell stage embryo. The distribution of injected morpholino oligonucleotides in the tadpole was monitored by using a fluorophore tagged to the three-prime end of the morpholino oligonucleotide. Anti-xTrkB morpholino oligonucleotides injected into right or left dorsal blastomeres were distributed in the right or left eye and the right or left tectum of the tadpole, respectively. The specifically reduced expression of TrkB was confirmed by western blots of TrkB in tadpole brain sections (Q. Zhou, Z. R. Wang, S. T. Wong, H. W. Tao and M.M.P, unpublished data). Recordings

were made from the oligonucleotide positive retina, in which most RGCs showed strong fluorescence.

Electrophysiology and visual stimulation

Whole-cell perforated recording from RGCs or tectal cells were made under visual control by methods described previously²⁸. The micropipettes were made from borosilicate glass capillaries (Kimax), had a resistance in the range of 2–3 MΩ, and were tip-filled with internal solution and then were back-filled with internal solution containing 200 μg ml^{−1} amphotericin B. The internal solution contained: 115 mM K-gluconate, 4 mM NaCl, 1.5 mM MgCl₂, 20 mM HEPES at pH 7.3 and 0.5 mM EGTA. Recordings were made with a patch clamp amplifier (Axopatch 200A; Axon Instruments). The whole cell capacitance was fully compensated and the series resistance (10–20 MΩ) was compensated at 75–80% (lag 60 μs) and monitored during the experiment. The data were discarded if the series resistance varied by >20%. The input resistance of RGCs at −60 to −70 mV was usually in the range of 0.5 to 1 GΩ. Signals were filtered at 5 kHz and sampled at 10 kHz using Axoscope software (Axon Instruments). For cell-attached recording, the pipette was filled with bath solution and a loose seal was made. To record mEPSCs from RGCs or tectal neurons, TTX (1 μM) was added to external solution. A total amount of 400 ng BDNF (PeproTech), dissolved in 15 μl bath solution containing 0.1% bovine serum albumin (BSA), was puffed into the third ventricle near the tectum through a Picospritzer II (General Valve), using a micropipette with a tip opening of 6 μm and repetitive pressure pulses (5 psi, 0.5 Hz, 1 s duration). It usually took 15 min for the BDNF-containing solution to be completely ejected. A small LCD-screen (Sony, PLM-A35) was mounted on the camera port of the microscope, allowing projection of computer-generated images onto the retina of the tadpole. Light responses were evoked by whole-field stimulations with a step (4 s) increase in the whole-field light illumination. All drugs were from Sigma.

Data analysis

Mini analysis (6.0.1 from Synaptosoft) was used to detect and analyse mEPSCs, sEPSCs and spikes, and the same program was used for peak-scaled non-stationary noise analysis of mEPSCs or sEPSCs, using the method previously described^{29,30}. The relationship between the variance (σ^2) and the amplitude (I) of mEPSCs or sEPSCs was fitted by the parabola $\sigma^2 = i \times I - I^2/N$, where the estimated weighted single-channel current (i) and the mean receptor number per synapse (N) were derived from the best fit. The estimated weighted single-channel conductance (g_s), was derived from the relationship: $g_s = i/(E_{rev} - V_c)$, where E_{rev} is the reversal potential of EPSCs, and V_c the holding potential. The mean channel open probability of synaptic receptors (p_o), was calculated by the function: $p_o = I_m/(i \times N)$, where I_m is the amplitude of mean mEPSCs or sEPSCs. A prerequisite of applying the noise analysis is that data are not filtered by the recording system and electrotonic structure of the cell (see Supplementary Methods). Statistical analysis was performed using paired Student's t -test and summary data are presented as means $\pm$ s.e.m.

Received 17 December 2003; accepted 7 May 2004; doi:10.1038/nature02618.

1. Fitzsimonds, R. M., Song, H. J. & Poo, M. M. Propagation of activity-dependent synaptic depression in simple neural networks. *Nature* **388**, 439–448 (1997).
2. Tao, H. W., Zhang, L. I., Bi, G. Q. & Poo, M. M. Selective presynaptic propagation of long-term potentiation in defined neural networks. *J. Neurosci.* **20**, 3233–3243 (2000).
3. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagation errors. *Nature* **323**, 533–536 (1986).
4. Cohen-cory, S. & Fraser, S. E. BDNF in the development of the visual system of *Xenopus*. *Neuron* **12**, 747–761 (1994).
5. Huang, E. J. & Reichardt, L. F. Neurotrophins: roles in neuronal development and function. *Annu. Rev. Neurosci.* **24**, 677–736 (2001).
6. Lohof, A. M. & Poo, M. M. Potentiation of developing neuromuscular synapses by the neurotrophins NT3 and BDNF. *Nature* **363**, 350–353 (1993).
7. Lessmann, V., Gottmann, K. & Heumann, R. BDNF and NT-4/5 enhance glutamatergic synaptic transmission in cultured hippocampal neurons. *Neuroreport* **6**, 21–25 (1994).
8. Li, Y. X. *et al.* Expression of a dominant negative TrkB receptor, T1, reveals a requirement for presynaptic signaling in BDNF-induced synaptic potentiation in cultured hippocampal neurons. *Proc. Natl Acad. Sci. USA* **95**, 10884–10889 (1998).
9. Schinder, A. F., Berninger, B. & Poo, M. M. Postsynaptic target specificity of neurotrophin-induced presynaptic potentiation. *Neuron* **25**, 151–163 (2000).
10. Poo, M. M. Neurotrophins as synaptic modulators. *Nature Rev. Neurosci.* **2**, 24–32 (2001).
11. Isaac, J. T., Nicoll, R. A. & Malenka, R. C. Evidence for silent synapses: implications for the expression of LTP. *Neuron* **15**, 427–434 (1995).
12. Durand, G. M., Kovalchuk, Y. & Konnerth, A. Long-term potentiation and functional synapse induction in developing hippocampus. *Nature* **381**, 71–75 (1996).
13. Wu, G. Y., Malinow, R. & Cline, H. T. Maturation of a central glutamatergic synapse. *Science* **274**, 972–976 (1996).
14. Watson, F. L. *et al.* Rapid nuclear responses to target-derived neurotrophins require retrograde transport of ligand-receptor complex. *J. Neurosci.* **19**, 7889–7900 (1999).
15. von Bartheld, C. S., Byers, M. R., Williams, R. & Bothwell, M. Anterograde transport of neurotrophins and axodendritic transfer in the developing visual system. *Nature* **379**, 830–833 (1996).
16. Ginty, D. D. & Segal, R. A. Retrograde neurotrophin signaling: Trk-ing along the axon. *Curr. Opin. Neurobiol.* **12**, 268–274 (2002).
17. Purves, D. Functional and structural changes in mammalian sympathetic neurons following interruption of their axons. *J. Physiol. (Lond.)* **252**, 429–463 (1975).
18. Purves, D. Effects of nerve growth factor on synaptic depression after axotomy. *Nature* **260**, 535–536 (1976).
19. Wong, W. T., Faulkner-Jones, B. E., Sanes, J. R. & Wong, R. O. Rapid dendritic remodeling in the developing retina: dependence on neurotransmission and reciprocal regulation by Rac and Rho. *J. Neurosci.* **20**, 5024–5036 (2000).

20. Lom, B., Cogen, J., Sanchez, A. L., Vu, T. & Cohen-cory, S. Local and target-derived brain-derived neurotrophic factor exert opposing effects on the dendritic arborization of retinal ganglion cells *in vivo*. *J. Neurosci.* **22**, 7639–7649 (2002).
21. Hartmann, M., Heumann, R. & Lessmann, V. Synaptic secretion of BDNF after high-frequency stimulation of glutamatergic synapses. *EMBO J.* **20**, 5887–5897 (2001).
22. Balkowiec, A. & Katz, D. M. Cellular mechanisms regulating activity-dependent release of native brain-derived neurotrophic factor from hippocampal neurons. *J. Neurosci.* **22**, 10399–10407 (2002).
23. Chytrova, G. & Johnson, J. E. Spontaneous retinal activity modulates BDNF trafficking in the developing chick visual system. *Mol. Cell. Neurosci.* **25**, 549–557 (2004).
24. Constantine-Paton, M., Cline, H. T. & Debski, E. Patterned activity, synaptic convergence, and the NMDA receptor in developing visual pathways. *Annu. Rev. Neurosci.* **13**, 129–154 (1990).
25. Wingate, R. J. & Thompson, I. D. Targeting and activity-related dendritic modification in mammalian retinal ganglion cells. *J. Neurosci.* **14**, 6621–6637 (1994).
26. Churchland, P. S. & Sejnowski, T. J. *The Computational Brain* (MIT Press, Cambridge, Massachusetts, 1992).
27. Nieuwkoop, P. D. & Faber, J. *Normal table of Xenopus laevis*, 2nd edn (North Holland, Amsterdam, 1967).
28. Zhang, L. I., Tao, H. W., Holt, C. E., Harris, W. A. & Poo, M. M. A critical window for cooperation and competition among developing retinotectal synapses. *Nature* **395**, 37–44 (1998).
29. Traynelis, S. F., Silver, R. A. & Cull-Candy, S. G. Estimated conductance of glutamate receptor channels activated during EPSCs at the cerebellar mossy fiber-granule cell synapse. *Neuron* **11**, 279–289 (1993).
30. Benke, T. A., Lüthi, A., Isaac, J. T. R. & Collingridge, G. L. Modulation of AMPA receptor unitary conductance by synaptic activity. *Nature* **393**, 793–797 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank X. H. Zhang for writing the software of light stimulation and Z. R. Wang for establishing the method of TrkB knockdown of *Xenopus* tadpoles. This work was supported by grants from the NIH.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.M.P. (mpoo@uclink.berkeley.edu).

Gene regulation and DNA damage in the ageing human brain

Tao Lu¹, Ying Pan¹, Shyan-Yuan Kao¹, Cheng Li², Isaac Kohane³, Jennifer Chan⁴ & Bruce A. Yankner¹

¹Department of Neurology and Division of Neuroscience, The Children's Hospital and Harvard Medical School, Enders 260,300 Longwood Avenue, Boston, Massachusetts 02115, USA

²Department of Biostatistics, Harvard School of Public Health, and ³Department of Medicine, The Children's Hospital and Harvard Medical School, and

⁴Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA

The ageing of the human brain is a cause of cognitive decline in the elderly and the major risk factor for Alzheimer's disease¹. The time in life when brain ageing begins is undefined^{2–4}. Here we show that transcriptional profiling of the human frontal cortex from individuals ranging from 26 to 106 years of age defines a set of genes with reduced expression after age 40. These genes play central roles in synaptic plasticity, vesicular transport and mitochondrial function. This is followed by induction of stress response, antioxidant and DNA repair genes. DNA damage is markedly increased in the promoters of genes with reduced expression in the aged cortex. Moreover, these gene promoters are selectively damaged by oxidative stress in cultured human neurons, and show reduced base-excision DNA repair. Thus, DNA damage may reduce the expression of selectively vulnerable genes involved in learning, memory and neuronal survival, initiating a programme of brain ageing that starts early in adult life.

To investigate age-dependent regulation of gene expression in the human brain, RNA was harvested from postmortem samples of the

frontal pole of 30 individuals ranging in age from 26 to 106 and was analysed using Affymetrix gene chips. To resolve genes with similar age-dependent expression patterns, the data were analysed for genes that correlate significantly with age and visualized by hierarchical clustering. This analysis demonstrated a cluster of co-regulated genes with reduced expression, and another cluster of genes with increased expression in aged individuals (Fig. 1a). To assess the rate of these gene changes, the entire transcriptome profile was compared at each age, and Pearson correlation coefficients were derived as a measure of similarity between any two ages (Fig. 1b). The group of individuals ≤ 42 years old showed the most homogeneous pattern of gene expression, and the group ≥ 73 years old was also relatively homogeneous (red colour indicating positive correlation).

Moreover, these two age groups were negatively correlated with each other (blue colour indicating negative correlation). In contrast, the middle age group ranging in age from 45–71 exhibited much greater heterogeneity, with some cases resembling the young group and others resembling the aged group (Fig. 1b). These results suggest that a genetic signature of human cortical ageing may be defined starting in young adult life, and that the rate of age-related change may be heterogeneous among middle age individuals.

Age-related genes were identified by performing statistical group comparison of frontal cortical samples from individuals ≤ 42 and ≥ 73 years old. About 4% of the approximately 11,000 genes analysed were significantly changed (1.5-fold or more, Supplementary Table 2). To validate the microarray data, we compared it with

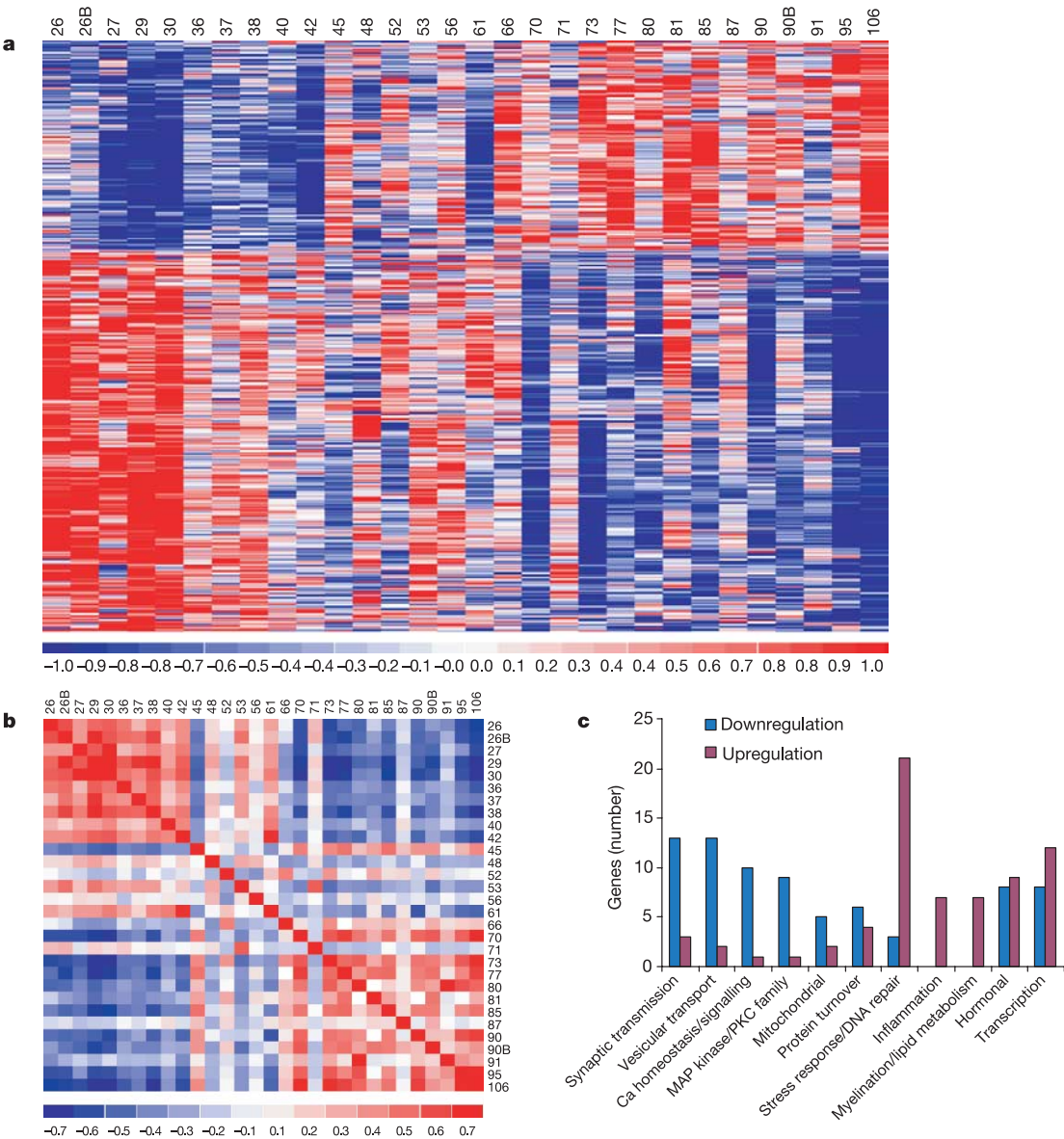


Figure 1 Ageing and gene expression in the human brain. **a**, The transcriptional profile of the normal human prefrontal cortex from 30 individuals ranging in age from 26–106 was analysed by hierarchical clustering of age-regulated genes. One cluster of genes exhibits reduced expression (transition from red to blue), and another cluster exhibits increased expression (transition from blue to red) in aged individuals. Age in years is shown above each lane. Standardized expression values of genes are displayed according to the colour scale, in which red represents above average expression and blue represents below

average expression. Absolute fold changes of individual genes are shown in Table 1 and Supplementary Table 2. **b**, Gene expression profiles were compared between all ages to derive a matrix of Pearson correlation coefficients that indicate the degree of overall similarity between any two cases (see Methods). Positive correlation is indicated by red and negative correlation by blue. Age in years of each case is indicated. The letter B is appended to ages for which there are two independent cases. **c**, Relative changes in gene ontology categories in the aged cortex.

quantitative real-time polymerase chain reaction (PCR) for a subset of functionally important genes. Microarray analysis and quantitative PCR generally showed consistent changes (Fig. 2a). Furthermore, consistent changes at the protein level were observed for a subset of genes analysed by western blotting (Fig. 2b). The post-mortem interval did not correlate significantly with the messenger RNA expression levels of 40 age-regulated genes examined, or with a cumulative measure of all the genes in each of the two age-related clusters (see 'statistical data analysis' in Supplementary Methods). In addition, expression of a number of neuron-specific markers, including β -tubulin, contactin 2 (TAG-1), GAP-43, γ -enolase, and syntaxin 1, did not change significantly with age, suggesting that ageing was not associated with major changes in neuronal cell number.

Genes that play a role in synaptic function and the plasticity that underlies learning and memory were among those most significantly affected in the ageing human cortex (Table 1, Figs 1c and 2). Several neurotransmitter receptors that are centrally involved in synaptic plasticity^{5,6} showed significantly reduced expression after age 40, including the GluR1 AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) receptor subunit, the NMDA (*N*-methyl-D-aspartate) R2A receptor subunit, and subunits of the GABA_A receptor. Moreover, the expression of genes that mediate synaptic vesicle release and recycling was significantly reduced, notably VAMP1/synaptobrevin, synapsin II, RAB3A and SNAPS.

Members of the major signal transduction systems that mediate long-term potentiation (LTP) and memory storage were age-down-regulated, notably the synaptic calcium signalling system, with reduced expression of calmodulin 1 and CAM kinase II α (Table 1 and Fig. 2a, b). The major calcium-binding proteins calbindins 1 and 2, the calcium pump ATP2B2, and the calcium-activated transcription factor MEF2C that promotes neuronal survival^{7,8}, were also significantly reduced. Furthermore, multiple members of the protein kinase C (PKC) and Ras-MAP (mitogen-activated

protein) kinase signalling pathways showed decreased expression. The activation state of PKC was also reduced, as indicated by decreased levels of activated phosphorylated forms (Fig. 2b). Thus, calcium homeostasis and neuronal signalling may be affected in the aged cortex.

Genes involved in vesicular/protein transport showed reduced expression in the aged cortex, including multiple RAB GTPases, sortilin, dynein, and clathrin light chain (Table 1). Moreover, microtubule-associated proteins (MAP1B, MAP2, tau and kinesin 1B) that stabilize microtubules and promote axonal transport were consistently and robustly reduced. The p35 activator of cyclin-dependent kinase-5 (cdk5), which regulates intraneuronal protein trafficking and synaptic function⁹, was also significantly reduced. Thus, vesicular trafficking may be affected in the aged human cortex. In addition, a number of genes involved in protein turnover also showed reduced expression in aged cortex, including ubiquitin-conjugating enzymes, the lysosomal proton pump, and the enzymes D-aspartate O-methyltransferase and methionine adenosyltransferase II, which repair damaged proteins.

The ageing of the human frontal cortex was also associated with increased expression of genes that mediate stress responses and repair (Fig. 1c and Table 1). These included genes involved in protein folding (heat shock protein 70 and α crystallin), antioxidant defence (nonselenium glutathione peroxidase, paraoxonase and selenoprotein P) and metal ion homeostasis (metallothioneins 1B, 1G and 2A). Genes involved in inflammatory or immune responses, such as tumour-necrosis factor (TNF)- α , were also increased. Increased expression of the base-excision repair enzymes 8-oxoguanine DNA glycosylase and uracil DNA glycosylase is consistent with increased oxidative DNA damage in the aged cortex.

The pronounced downregulation of a defined gene cluster followed by induction of antioxidant and DNA repair genes led us to hypothesize that oxidative DNA damage might target specific genes. Promoter regions may be especially vulnerable, as they contain (G + C)-rich sequences that are highly sensitive to oxidative DNA damage and are not protected by transcription-coupled repair¹⁰. To explore this hypothesis, we devised an assay to detect DNA damage in specific gene sequences. Genomic DNA was isolated under conditions that prevent *in vitro* oxidation, and then cleaved with formamidopyrimidine-DNA glycosylase (FPG), which is an N-glycosidase and AP-lyase that selectively releases damaged bases from DNA, predominantly affecting the major oxidation product 8-oxoguanine¹¹. FPG creates a single-strand break at the apurinic site, rendering it resistant to PCR amplification. Hence, DNA damage to specific sequences can be determined from the ratio of intact PCR products in cleaved versus uncleaved DNA using quantitative PCR. We initially used this assay to assess damage in genomic DNA from fetal human cortex. Fetal cortical DNA did not show significant oxidative DNA damage in the 1-kb upstream promoter regions of several genes that show age-related changes in expression in the adult brain (Fig. 3a).

We then determined whether DNA damage increases in the ageing human cortex, and whether there is a predilection for specific genes. To address this question, we examined the promoters of 30 different genes. Each of these genes showed increased promoter DNA damage in the aged cortex. DNA damage appeared in many genes after age 40, and was most pronounced in all genes after age 70 (Fig. 3b, c). DNA damage also occurred in the exons of these genes with a similar time course, but to a lesser extent than in the promoter regions (data not shown). Biopsy samples of human cortex from individuals who underwent elective neurosurgical procedures showed a similar pattern of age-related DNA damage as postmortem samples (Fig. 3c, asterisks). Thus, DNA damage is pervasive in the ageing human cortex.

We then asked whether gene downregulation in the ageing brain is associated with accelerated DNA damage. To address this question, we determined the increase in promoter DNA damage,

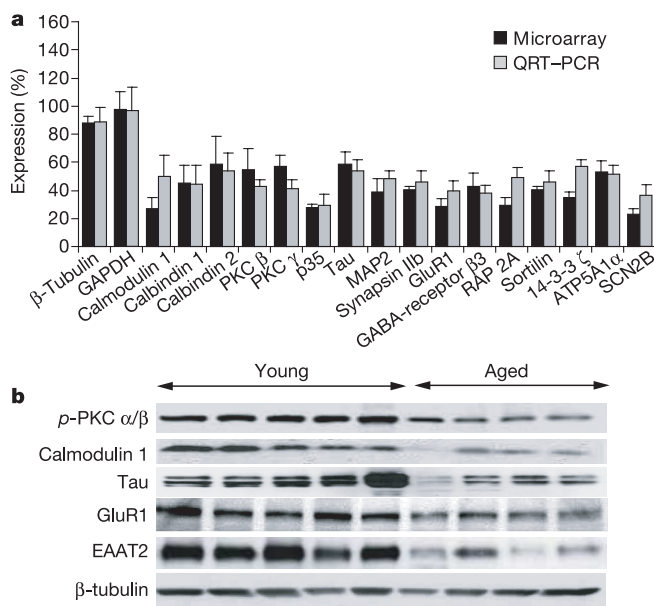


Figure 2 Confirmation of microarray results for synaptic, calcium homeostasis and transport-related genes. **a**, Shown are mRNA levels of selected genes in the aged frontal cortex determined by microarray analysis and quantitative RT-PCR. Values are percentage mRNA levels in aged cases (≥ 73 years old) versus young cases (≤ 42 years old) and represent the mean $\pm$ s.d.; $n = 4$. **b**, Age-related protein levels. Shown are immunoblots from five young and four aged frontal cortical samples. The p-PKC α/β blot specifically resolves activated phosphorylated forms of PKC α/β . EAAT2 is the predominant human brain glutamate transporter.

Table 1 **Age-regulated genes in the human frontal cortex**

Function	Gene name	Accession number	Fold Δ	q value
Synaptic function				
Synaptic transmission	GluR1	M81886	-2.2 to -2.4	0.002
	NMDA receptor 2A	U09002	-2.3	0.002
	GABA A receptor β 3	M82919	-3.2	0.002
	GABA A receptor α	AF016917	-1.5	0.002
	Serotonin receptor 2A	AA418537	-2.0	0.002
	Voltage-gated Na channel II β (SCN2B)	AF049498	-5.1	0.002
	Voltage-dependent calcium channel β 2	U95019	-1.9	0.002
	Neurexin 1	AB011150	-1.6	0.002
	Synaptobrevin 1 (VAMP1)	M36200	-3.4	0.002
	Synapsin II b	U40215	-3.4	0.002
	γ SNAP	U78107	-2.2	0.002
	α SNAP	U39412	-1.6	0.005
	RAB3A	M28210	-1.7	0.002
	SNAP23	AJ011915	1.7	0.005
	Synaptophysin-like protein	X68194	1.8	0.006
	Calmodulin 1	U12022	-2.2 to -4.1	0.002
	Calmodulin 3	J04046	-1.6	0.002
	Calbindin 1 (28 kD)	AF068862	-2.5	0.002
	Calbindin 2 (29 kD, calretinin)	X56667	-1.6	0.003
	CaM kinase II α	AB023185	-1.7	0.008
Ca ²⁺ homeostasis/signalling	CaM kinase IV	D30742	-2.0	0.007
	Calcineurin B α	M30773	-2.8	0.002
	ATPase, Ca ²⁺ -transporting, plasma membrane 2 (ATP2B2)	L20977	-2.5	0.002
	ATPase, Ca ²⁺ -transporting, plasma membrane 2 (ATP2A2)	M23114	-1.6	0.002
	Regucalcin (senescence marker protein)	D31815	1.7	0.002
	Phosphodiesterase 4D	U02882	-1.9	0.002
	Adenylyl cyclase associated protein 2	HG2530	-1.6 to -2.3	0.002-0.003
	PKC β 1	X06318	-1.9 to -2.9	0.002
	PKC γ	Z15114	-1.8	0.002
	PKC ζ	Z15108	-1.7	0.002
G protein signalling	Rap2A	X12534	-3.8 to -4.1	0.002
	Regulator of G protein signalling 4	U27768	-1.8 to -2.2	0.002
MAP kinase cascades	G protein, q polypeptide (GNAQ)	U43083	-2.0	0.002
	MAPK1	Z11695	-1.9	0.002
	MAPK9	U09759	-1.7	0.003
	MAPKK4	U17743	-3.1	0.002
	Ras-GNRF	HG2510	-2.4 to -4.7	0.003-0.008
	MAPKK5	U67156	1.6	0.002
	14-3-3 ζ	U28964	-3.6	0.002
	p21 activated protein kinase (PAK1)	U24152	-2.7	0.002
	Cdk5, regulatory subunit 1 (p35)	X80343	-3.4	0.002
	Cdk5			
Vesicular transport				
	RAB1A	M28209	-1.6	0.006
	RAB3A	M28210	-1.7	0.002
	RAB5A	M28215	-1.9	0.002
	RAB6A	M28212	-3.5	0.002
	Kinesin 1B	AB011163	-2.2	0.002
	Sortilin 1	X98248	-3.5	0.002
	Dynein (DNCH1)	H05552	-2.4	0.002
	Dynamin 1-like	AF000430	-1.6	0.002
	Trans Golgi network protein 2	AF027516	-2.2	0.002
	Golgi reassembly stacking protein 2	W26854	-1.7	0.002
	Phosphatidylinositol transfer protein β	D30037	-2.0	0.002
	Clathrin, light polypeptide	M20470	-1.6	0.002
	Kinesin 2	L04733	1.7	0.005
	VAMP3	H93123	1.5	0.002
	MAP1B	L06237	-4.9	0.002
	MAP2	U01828	-2.1 to -4.2	0.002
	Tau	J03778	-2.3	0.002
	RAN binding protein 9	AF064606	-1.7	0.005
Neuronal survival				
	MADS box transcription enhancer factor 2C (MEF2C)	S57212	-2.7	0.002
	Inositol polyphosphate-4-phosphatase I	AI955897	-2.0	0.002
	Inositol 1,4,5 trisphosphate 3 kinase A	X54938	-2.5	0.002
	Inositol 1,4,5 trisphosphate 3-kinase B	X57206	1.9	0.002
Protein turnover				
	ATPase, H ⁺ -transporting, lysosomal V1 subunit H	W27838	-2.5	0.005
	ATPase, H ⁺ -transporting, lysosomal V1 subunit A	L09235	-1.7	0.007
	ATPase, H ⁺ -transporting, lysosomal V1 subunit G 2	W26326	-1.5	0.002
	Ubiquitin conjugating enzyme Ubch5	HG3344	-1.6	0.002
	Ubiquitin conjugating enzyme E2M	AF075599	-1.6	0.002
	Ubiquitin carrier protein	M91670	-1.6	0.002
	Lysosomal associated membrane protein 2	U36336	2.3	0.002
	Calpastatin (calpain inhibitor)	D16217	1.6	0.007
	Serine/cysteine proteinase inhibitor	D83174	1.7	0.005
	Angiotensinogen (serine /cysteine) proteinase Inhibitor A8	K02215	1.6	0.002
Amino acid modification				
	Protein-L-isoaspartate (Daspertate) O-methyltransferase	D25547	-2.7	0.002
	Methionine adenosyltransferase II α	X68836	-2.1	0.002
	Beta-1,3-galactosyltransferase	Y15062	-1.9	0.002
	Glutamate decarboxylase 1	M81883	-1.6	0.005
	Methionine synthase reductase	AF025794	1.6	0.008

Table 1 – continued

Function	Gene name	Accession number	Fold Δ	q value
Mitochondrial	Transglutaminase 2	M55153	2.8	0.002
	Glycine amidinotransferase	S68805	1.5 to 1.8	0.002
	Lysine hydroxylase 2	U84573	2.4	0.002
	ATP synthase, H ⁺ -transporting, mitochondrial F1α1	D14710	−2.3	0.002
	Mitochondrial ribosomal protein L28	U19796	−1.7	0.002
	Mitochondrial ribosomal protein S12	Y11681	−2.2	0.002
	Cytochrome c synthase	U36787	−1.6	0.002
	Translocase of inner mitochondrial membrane 17 A	X97544	−2.0	0.002
Stress response	Monoamine oxidase A	AA420624	1.6	0.002
	Mitochondrial 3-oxoacyl-Coenzyme A thiolase	D16294	1.5	0.003
Antioxidant	Nonselenium glutathione peroxidase	D14662	1.7	0.002
	Selenoprotein P	Z11793	1.7	0.002
DNA repair	Paraoxonase 2	AF001601	1.6	0.002
	Cystathionine-beta-synthase	L00972	1.6	0.002
	8-oxoguanine DNA glycosylase	U88620	1.6	0.006
	Uracil-DNA glycosylase	Y09008	1.7	0.005
Stress	Topoisomerase I binding protein	U82939	1.6	0.009
	Topoisomerase II β	M27504	−1.7	0.003
	FK506 binding protein 12-rapamycin associated protein 1	L34075	−1.9	0.002
	Heat shock 70 kD protein 2	L26336	1.9 to 2.2	0.005–0.006
Metal ion homeostasis	Crystallin, alpha B	AL038340	1.6 to 2.0	0.002–0.003
	Hypoxia inducible factor 1 α (HIF1 α)	U22431	2.0	0.005
	HIF-1 responsive RTP801	AA522530	2.5	0.002
	Transglutaminase 2	M55153	2.8	0.002
	p53 binding protein 2	U58334	1.7	0.002
	Retinoblastoma-associated protein 140	AB029028	1.8	0.007
	Retinoblastoma-like 2 (p130)	X76061	1.6	0.007
	Stress 70 protein chaperone	U04735	−1.8	0.006
	Metallothionein 1G	J03910	2.2	0.005
	Metallothionein 1B	M13485	1.6	0.002
Inflammation	Metallothionein 2A	R92331	1.5 to 1.7	0.002
	Haem binding protein 2	W27949	2.0	0.002
	Haemoglobin β	L48215	2.9 to 3.3	0.002
	Hephaestin	AB014598	1.6	0.002
Myelination/lipid metabolism	TNF-α	AF010312	2.7	0.006
	C type lectin	X96719	2.7	0.003
	H factor (complement)-1	M65292	3.2	0.002
	Interferon, gamma-inducible protein 16	M63838	2.0	0.005
	Interferon regulatory factor 7	U53831	1.9	0.003
	Integrin α5	M14648	1.8	0.002
	Integrin β1	X07979	1.7	0.002
	Oligodendrocyte lineage transcription factor 2	U48250	1.7	0.002
Transcription	Peripheral myelin protein 22	D11428	1.7	0.003
	Proteolipid protein 1	M54927	1.6	0.002
	Fatty acid desaturase 1	AF009767	1.8 to 2.1	0.002
	Apolipoprotein D	J02611	2.1	0.002
	Low density lipoprotein receptor related protein 4	AB011540	2.0	0.002
	Sterol carrier protein 2	U11313	1.6	0.002
	Phospholipase D3	U60644	−1.6	0.002
	Transcription factor ZHX2	AB020661	2.1	0.002
Hormonal	NK2 transcription factor	AF019415	1.5 to 1.8	0.005–0.008
	Inhibitor of DNA binding 4 (ID4)	AL022726	2.3	0.002
	Zinc finger protein 238	U38896	−4.6	0.002
	Forkhead box G1A	X74143	−2.0	0.002
	Chromatin remodelling complex (SMARCC2)	D26155	−1.8	0.002
	ETS2	J04102	−1.7	0.002
	E2F transcription factor 4	S75174	−1.6	0.003
	Insulin receptor	X02160	1.6	0.004
	Leptin receptor	AW026535	1.7	0.002
	Orexin receptor	AF041245	1.6	0.002
	Vascular endothelial growth factor	AF022375	1.8	0.005
	Secreted frizzled related protein 1	AF056087	1.9	0.009
	FGF receptor 2	M87770	1.6	0.002
	FGF receptor 3	M64347	1.8	0.002
	FGF2 (basic)	J04513	2.1	0.002
	Proenkephalin	J00123	−2.5	0.003
	Somatostatin	AI636761	−1.8 to −2.9	0.002
	Cholecystokinin B receptor	L10822	−2.7	0.002
	Chromogranin B (secretogranin 1)	Y00064	−1.6	0.007
	RevErbA β receptor (NR1D2)	D16815	−2.9	0.003
	GDNF receptor α2	AF002700	−1.6	0.003
	FGF 12	AL119322	−2.4 to −2.6	0.002–0.008
	FGF 13	U66198	−2.3 to −3.2	0.002–0.005

Shown are selected age-regulated genes representative of functional groups. Age-downregulated genes are blue and age-upregulated genes are red. Fold changes and statistical *q* values with a range reflect multiple probe sets for the same gene. Gene accession numbers are provided. See Supplementary Table 2 for a complete list of age-regulated genes.

indicated by the reduction in intact DNA, in individuals over 70 years old. This index of DNA damage was compared for genes that were downregulated, upregulated or stably expressed in the aged cortex. Stably expressed and upregulated genes showed a narrow range of promoter DNA damage in aged cortex (Fig. 3d). In contrast, most of the age-downregulated genes showed significantly greater DNA damage in the aged cortex ($P < 0.001$) (Fig. 3d). These results were confirmed by independently assaying 8-oxoguanine through cleavage of genomic DNA with the 8-oxoguanine-specific

N-glycosylase human OGG1. 8-oxoguanine levels were markedly increased in the promoters of most of the age-downregulated genes examined (Fig. 3e). Chromatin immunoprecipitation of the calmodulin 1 promoter with a monoclonal antibody to 8-oxoguanine confirmed an approximately eightfold increase in 8-oxoguanine in aged cortical samples (Fig. 3f). Thus, accelerated DNA damage is associated with reduced gene expression in the aged human cortex.

To obtain greater insight into the effects of DNA damage on gene expression, we produced human neuroblastoma SH-SY5Y cell lines

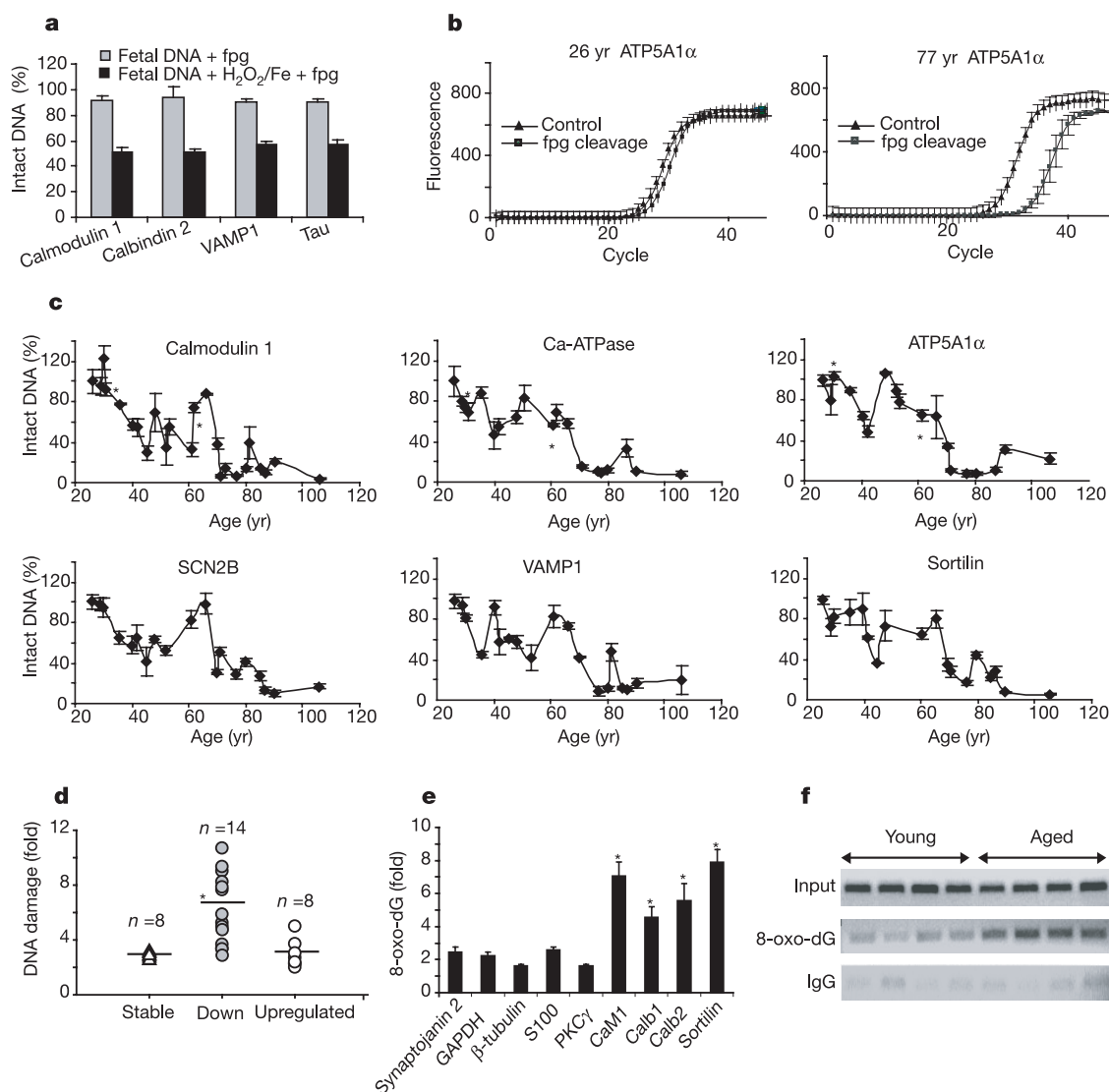


Figure 3 DNA damage in the ageing human cortex. **a**, Genomic DNA from fetal cortex does not exhibit significant DNA damage. DNA damage to the promoter regions of the indicated genes was assayed by cleavage with the endoglycosidase FPG and quantitative PCR. Intact DNA is the percentage detected by PCR following FPG cleavage relative to that in uncleaved DNA. **b**, Ageing increases oxidative DNA damage to the mitochondrial ATP synthase α (ATP5A1 α) promoter. Shown are real-time fluorescence PCR curves from 26- and 77-year-old frontal cortical samples. Note the marked shift in PCR cycle number following FPG cleavage of 77 yr old DNA. Values in **a** and **b** represent the mean $\pm$ s.d. **c**, Time course of DNA damage in the ageing frontal cortex. DNA damage was assayed in the promoters of age-downregulated genes (calmodulin 1, Ca-ATPase, ATP5A1 α , sodium channel β 3 (SCN2B), VAMP1, and sortilin) in cortical samples from 26- to 106-year-old cases and normalized to the 26-year-old value (100%). Values represent the mean $\pm$ s.d.; $n = 3$. Asterisks indicate intracortical biopsy samples. **d**, DNA damage to

promoters of genes that are stably expressed, downregulated or upregulated in the aged cortex. Shown is the fold increase in promoter DNA damage in aged cases (≥ 70 years old) relative to the youngest, 26-year-old case. Each point represents a gene (see 'DNA damage assay' in Supplementary Methods for gene identities). Asterisk indicates $P < 0.001$ relative to age-stable genes by analysis of variance (ANOVA) with post-hoc Student–Newman–Keuls test. **e**, Oxidative damage to gene promoters in the aged cortex. Shown is the fold increase in 8-oxoguanine (8-oxo-dG) incorporation into promoters of age-stable (GAPDH, β -tubulin and synaptotagmin 2), age-upregulated (S100), and age-downregulated genes (calmodulin 1 (CaM1), calbindin 1 (Calb1), calbindin 2 (Calb2), sortilin and PKC γ). Asterisks indicate $P < 0.05$ relative to GAPDH. Values represent the mean $\pm$ s.e.m.; $n = 4$. **f**, Chromatin immunoprecipitation of the calmodulin 1 promoter with a monoclonal antibody to 8-oxoguanine in aged (≥ 73 -year-old) and young (< 40 -year-old) cortical samples. Input DNA and non-specific IgG (IgG) controls are shown.

that stably overexpress the base-excision repair enzyme human OGG1^{12,13}. Oxidative DNA damage was induced by incubating SH-SY5Y cells with H₂O₂ and FeCl₂, resulting in rapid DNA damage to the promoter of the tau gene, followed by slow and incomplete DNA repair (Fig. 4a). In contrast, human OGG1-overexpressing

SH-SY5Y cells showed augmented DNA repair with complete restoration of intact DNA (Fig. 4a). Tau mRNA expression was also reduced by oxidative stress, but was completely restored by overexpression of human OGG1 (Fig. 4b). Cell viability was not significantly affected by the mild oxidative stress treatment or by

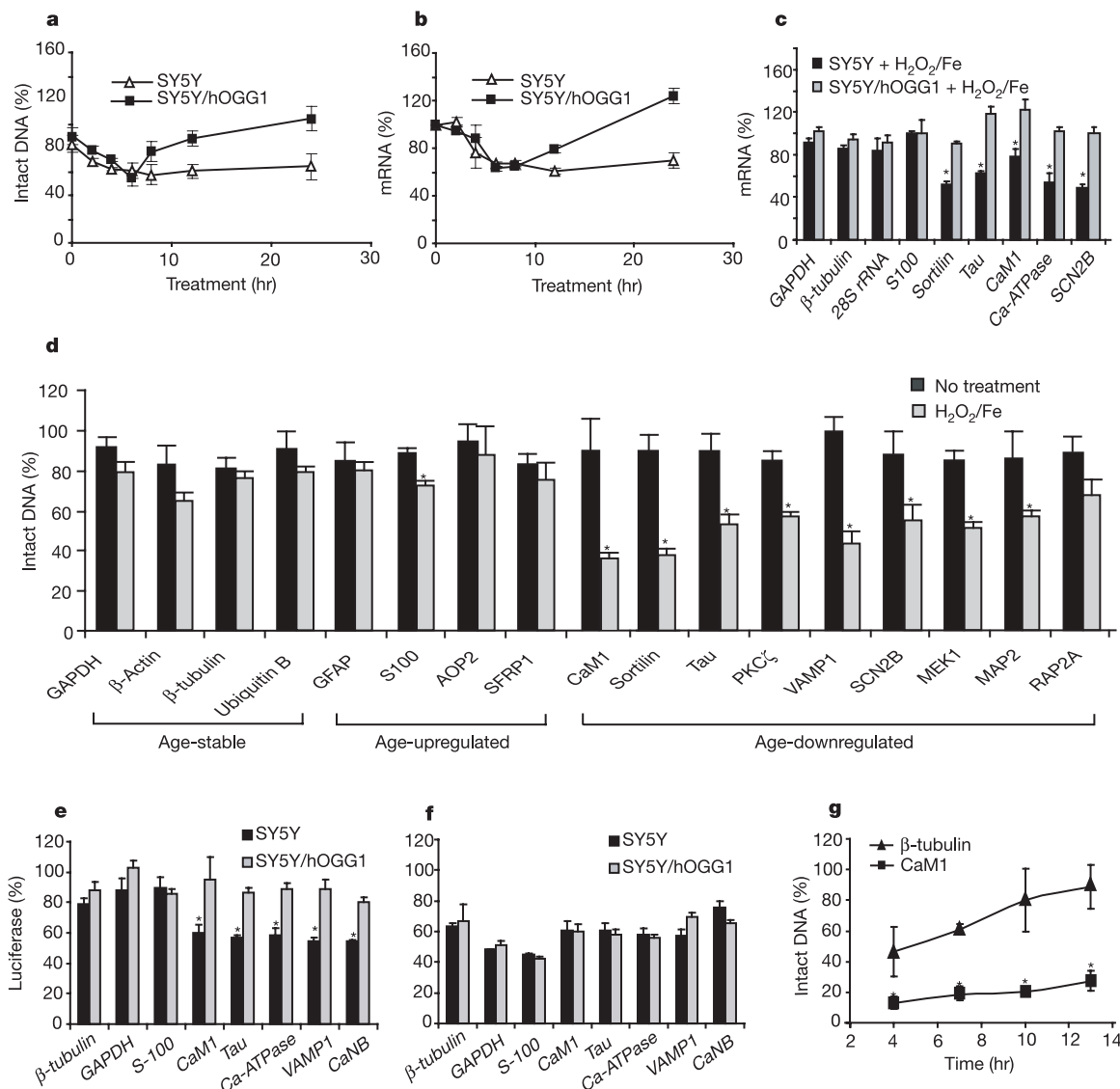


Figure 4 Promoters of age-downregulated genes show increased vulnerability to oxidative DNA damage. **a, b**, Human neuroblastoma SH-SY5Y cells were incubated with H₂O₂/FeCl₂ for the indicated time intervals to induce oxidative DNA damage. DNA damage (**a**) and mRNA expression (**b**) of the tau gene were determined in cells that overexpress the DNA repair enzyme human OGG1 (SY5Y/hOGG1) or the empty pcDNA3 vector (SY5Y). DNA damage was determined by the FPG cleavage/PCR-based assay. Values represent the mean $\pm$ s.e.m. **c**, mRNA levels of age-downregulated genes are selectively reduced by oxidative stress and restored by human OGG1. mRNA levels are expressed as percentage in the presence versus absence of H₂O₂/FeCl₂ and represent the mean $\pm$ s.e.m.; $n = 4$. Asterisk indicates $P < 0.05$ relative to no treatment by ANOVA with post-hoc Student–Newman–Keuls test. **d**, Increased vulnerability to oxidative DNA damage in promoters of age-downregulated genes. Human cortical neuronal cultures were incubated in the presence or absence of 100 μ M H₂O₂/20 μ M FeCl₂ for 12 hours and promoter DNA damage was assayed. Values represent the mean $\pm$ s.d.; $n = 3$. Asterisks indicate $P < 0.05$ relative to no treatment. $P < 0.001$ for the group of age-downregulated genes relative to age-stable or age-upregulated genes. **e**, Reduced

transcriptional activity of promoters of age-downregulated genes following oxidative DNA damage. Luciferase reporter plasmids derived from the promoters of age-downregulated genes (calmodulin 1 (CaM1), tau, Ca-ATPase, VAMP1/synaptobrevin, and calcineurin B (CaNB)) and genes without reduced expression (β -tubulin, GAPDH and S100) were incubated in the absence or presence of 100 μ M H₂O₂ for one hour *in vitro*, and then transfected into SH-SY5Y or SH-SY5Y/human OGG1 cells. Shown is luciferase activity of the damaged reporter expressed as percentage of the activity of the undamaged reporter after 16 h. Values represent the mean $\pm$ s.d.; $n = 4$. Asterisk indicates $P < 0.05$ for SH-SY5Y relative to SH-SY5Y/human OGG1. **f**, Ultraviolet damage does not discriminate between promoters of age-stable and age-downregulated genes. Values represent the mean $\pm$ s.d. **g**, DNA damage and repair of the β -tubulin and calmodulin 1 (CaM1) promoters. Reporter plasmids damaged *in vitro* by H₂O₂ were transfected and DNA damage was determined within each promoter sequence at increasing time intervals. Values are expressed relative to the transfected undamaged reporter, and represent the mean $\pm$ s.d.; $n = 3$. Asterisks indicate $P < 0.05$ relative to β -tubulin.

overexpression of human OGG1 (Supplementary Fig. 1). Thus, oxidative DNA damage can reduce gene expression.

Endogenous mRNA levels of a number of age-downregulated genes (tau, calmodulin 1, Ca-ATPase, sortilin and the sodium channel 2B) were significantly reduced by mild oxidative stress in SH-SY5Y cells, and restored by human OGG1 (Fig. 4c). In contrast, mRNA levels of genes that are not reduced in the ageing cortex (β -tubulin, GAPDH, S100 and 28S RNA) were not significantly affected (Fig. 4c). Thus, mRNA levels of some age-downregulated genes are highly sensitive to oxidative DNA damage.

We then surveyed a larger number of promoters from age-downregulated and age-stable genes to assess vulnerability to DNA damage in cultured human neurons. After pro-oxidative stress, the promoters of four age-stable and four age-upregulated genes showed minimal declines in the level of intact DNA (Fig. 4d). In contrast, eight of nine age-downregulated promoters showed significantly increased DNA damage. Thus, promoters of age-downregulated genes show increased vulnerability to oxidative DNA damage.

To determine whether reduced DNA repair contributes to promoter vulnerability, we cloned the promoters in luciferase reporter plasmids and performed a host cell reactivation assay¹⁴. Promoter reporter plasmids were damaged *in vitro* by either treatment with H₂O₂ or exposure to ultraviolet light, and then transfected into SH-SY5Y cells, together with an undamaged renilla luciferase control plasmid. Activation of H₂O₂-damaged reporters, an indicator of base-excision repair, was significantly reduced for reporters derived from the promoters of age-downregulated genes relative to reporters derived from age-stable genes (Fig. 4e). Reporter activity was restored by the base-excision repair enzyme human OGG1 (Fig. 4e). In contrast, activation of ultraviolet-damaged reporters was not significantly different for the two promoter categories, and was not affected by human OGG1 (Fig. 4f). Differential promoter damage was confirmed using the FPG cleavage/PCR-based assay. The calmodulin 1 promoter showed more DNA damage than the β -tubulin promoter, and was repaired very slowly (Fig. 4g). In contrast, the β -tubulin promoter was repaired much more rapidly ($P = 0.007$) (Fig. 4g). Thus, both increased initial damage and reduced base-excision repair may contribute to oxidative DNA damage in age-downregulated genes.

One factor that may predispose to DNA damage in the aged cortex is impaired mitochondrial function. Expression of the α subunit of the mitochondrial F1 ATP synthase, which couples oxidative phosphorylation to ATP synthesis, was significantly reduced in the aged human cortex (Table 1 and Fig. 2a). siRNA was used to reduce expression of F1 ATP synthase α mRNA and protein by 2.5-fold in SH-SY5Y cells (Supplementary Fig. 2a,b), approximating the reduction detected in aged human cortex. This resulted in a $24 \pm 1\%$ reduction in cellular ATP levels (mean $\pm$ s.d.; $n = 12$; $P = 0.014$), but did not affect overall cell viability or induce apoptosis. ATP synthase α small interfering RNA significantly increased promoter DNA damage in age-downregulated genes, and reduced mRNA levels (Supplementary Fig. 2c, d). Another siRNA (topoisomerase II β) and random 21-mer oligonucleotide controls had no significant effects (Supplementary Fig. 2a–d). DNA damage induced by knockdown of F1 ATP synthase α was partially reversed by the antioxidant vitamin E (Supplementary Fig. 2c, d). Thus, impaired mitochondrial function can lead to nuclear DNA damage.

Taken together, these findings suggest that accelerated DNA damage may contribute to reduced gene expression in the human brain after age 40. The cluster of age-downregulated genes includes many genes that play integral roles in synaptic plasticity^{5,6}, including NMDA and AMPA receptor function, calcium-mediated signalling, and synaptic vesicle release and recycling. In addition, the reduced expression of key calcium-binding and homeostatic genes in the aged cortex could compromise intraneuronal calcium homeostasis,

as observed in studies of ageing rodent neurons¹⁵, and may increase neuronal vulnerability to toxic insults. Our findings also provide support for the concept of ongoing oxidative stress in the ageing human cortex^{16–18}, as a variety of oxidative stress response and repair genes were induced. Similar stress response genes are induced in the ageing mouse and rat brain^{2–4}, and modulate ageing in *C. elegans*¹⁹. Thus, genome damage may compromise systems that subserve synaptic function and neuronal survival, leading to compensatory stress responses in the aged cortex.

Selective DNA damage to gene promoter sequences is a potential mechanism whereby the expression of specific genes could register the passage of time. Vulnerable DNA sequences appear to show increased initial DNA damage as well as reduced base excision repair. It will be of interest to define the specific vulnerable sequence motifs and their target transcription factors^{20,21}. A previous study showed that there is also substantial oxidative damage to mitochondrial DNA in the ageing human brain²². Our findings suggest that impaired mitochondrial function may contribute to the damage of vulnerable genes in the ageing cortex by increasing reactive oxygen species or by reducing ATP required for DNA repair.

When ageing begins and what triggers its onset is one of the major conundrums of biology. The young adult and extreme aged human populations are relatively homogeneous in their gene expression patterns in prefrontal cortex. However, the middle age population between 40 and 70 years of age exhibits much greater heterogeneity. Thus, individuals may diverge in their rates of ageing as they transit through middle age, approaching a state of 'old age' at different rates. It will be of interest to investigate this relationship in different populations and demographic groups. A recent report suggests that ageing in *C. elegans* and *Drosophila* is characterized by similar changes in orthologous mitochondrial and DNA repair genes, and that the pattern is established in early adulthood²³. Thus, measures to protect the genome early in adult life may influence the rate of subsequent functional decline and the vulnerability of the brain to age-related neurodegenerative diseases. □

Methods

DNA microarray analysis

Thirty cases spanning ages of 26 to 106 were used for microarray analysis (see Supplementary Information). Dissections of the frontal pole were performed and tissue samples were snap frozen in liquid nitrogen. Total RNA was extracted and complementary RNA targets were prepared, labelled and hybridized with an Affymetrix Test 3 Array. Samples with acceptable RNA quality (see Supplementary Information) were hybridized to Affymetrix HG-U95Av2 oligonucleotide arrays representing about 12,000 probe sets. Three approaches were used to analyse the data. (1) Arrays were normalized and genes that correlated with age (Spearman rank correlation P -value < 0.005) were determined and resolved by hierarchical clustering using dChip V1.3 software (see Supplementary Methods). (2) Correlation coefficient analysis was performed to assess the relatedness of each case to every other case using S-PLUS 2000 software (Insightful Corp.). Gene-wise standardized expression values of the genes that show Spearman rank correlation with age were used to compute Pearson correlation coefficients between two cases. The correlation coefficient matrix containing all pairwise correlation coefficients was then read into dChip for heat-map visualization. The range of observed correlation coefficients was -0.77 to 0.80 , and 0.7 was used as the display range (correlation above 0.7 is pure red, below -0.7 is pure blue, and 0 is white). (3) Significance analysis of microarrays (SAM) software²⁴ was used to compare young (≤ 42 years old) and aged (≥ 73 years old) groups to determine the list of genes with a ≥ 1.5 -fold change and median false discovery rate (FDR) < 0.01 (see Supplementary Tables 2 and 3). Some of the DNA microarray results were validated by quantitative reverse transcription (RT)-PCR and western blot analysis (see Supplementary Methods).

DNA damage analysis

The isolation of genomic DNA for the analysis of oxidative DNA damage was performed under conditions that prevent *in vitro* oxidation, including the presence of $50 \mu\text{M}$ of the free radical spin trap phenyl-tert-butyl nitron (PBN, Sigma), nitrogenation of all buffers, and avoidance of phenol and high temperature (see Supplementary Information). Fetal brain genomic DNA (18 weeks gestation) isolated under these conditions did not show significant oxidative damage (Fig. 3a). DNA damage was assayed by cleavage of genomic DNA with FPG (New England Biolabs), which acts as an efficient N-glycosylase and AP-lyase to excise 8-oxoguanine and other damaged bases, and creates a single-strand break that prevents PCR amplification. Quantitative RT-PCR was then used to determine the content of specific intact sequences. The ratio of PCR products after FPG cleavage to those present in uncleaved DNA was used to determine the percentage of intact DNA.

Incorporation of 8-oxoguanine was assayed by cleavage of genomic DNA with the 8-oxoguanine-specific N-glycosylase human OGG1 (New England Biolabs) and by chromatin immunoprecipitation with a monoclonal antibody to 8-oxoguanine (see Supplementary Methods).

Cell culture

Stable cell lines were derived from human neuroblastoma SH-SY5Y cells by transfection with a pcDNA3.1 vector encoding His-tagged nuclear human OGG1 (a gift from G. Verdine). Stable clonal cell lines were derived by selection in medium containing G418, and human OGG1 expression was confirmed by western blotting. Cells stably expressing the empty pcDNA3.1 vector were used as controls. Human cortical neuronal cultures were established as previously described²⁵. Cells were subjected to a mild pro-oxidative stress by treatment with H₂O₂ and FeCl₂, which did not reduce cell viability during the treatment period (see Supplementary Methods and Supplementary Fig. 1).

Luciferase reporter assays

Gene promoter sequences were identified based on published literature or predictions from the genome data base (see Supplementary Information), PCR-amplified from human brain genomic DNA, and cloned in the luciferase reporter vector pGL3-basic (Promega). The host cell reactivation assay of DNA repair¹⁴ was performed by treating luciferase reporter plasmids with 100 μ M H₂O₂ for one hour *in vitro*, or by exposing the DNA to ultraviolet-C light (254 nm) at 200 J m⁻². Damaged or control non-damaged promoter reporter plasmids were transfected into SH-SY5Y or SH-SY5Y/human OGG1 cells together with a pRL-TK-Renilla control plasmid (Promega) using Lipofectamine 2000 (Invitrogen). Sixteen hours after transfection, cells were lysed and analysed by the Dual-Luciferase Reporter Assay (Promega). Reporter luciferase activity was normalized to renilla-luciferase activity to control for transfection efficiency. The luciferase activity of H₂O₂ or ultraviolet-damaged reporters was expressed as the percentage of the luciferase activity of the corresponding non-damaged reporters. To assess DNA damage in the promoter regions of the transfected reporters, the FPG cleavage/PCR-based assay was used with PCR primers against regions of the pGL3 plasmid that encompassed the cloned promoters. This excluded amplification of endogenous promoter sequences of the target genes.

Received 5 March; accepted 14 May 2004; doi:10.1038/nature02661.

Published online 9 June 2004.

- Yankner, B. A. A century of cognitive decline. *Nature* **404**, 125 (2000).
- Lee, C.-K., Weindrich, R. & Prolla, T. A. Gene expression profile of the aging brain in mice. *Nature Genet.* **25**, 294–297 (2000).
- Jiang, C. H., Tsien, J. Z., Schultz, P. G. & Hu, Y. The effects of aging on gene expression in the hypothalamus and cortex of mice. *Proc. Natl Acad. Sci. USA* **98**, 1930–1934 (2001).
- Blalock, E. M. *et al.* Gene microarrays in hippocampal aging: statistical profiling identifies novel processes correlated with cognitive impairment. *J. Neurosci.* **23**, 3807–3819 (2003).
- Kandel, E. R. The molecular biology of memory storage: a dialog between genes and synapses. *Science* **294**, 1030–1038 (2001).
- Malinow, R. & Malenka, R. C. AMPA receptor trafficking and synaptic plasticity. *Annu. Rev. Neurosci.* **25**, 103–126 (2002).
- Mao, Z., Bonni, A., Xia, F., Nadal-Vicens, M. & Greenberg, M. E. Neuronal activity dependent cell survival mediated by transcription factor MEF2. *Science* **286**, 785–790 (1999).
- Okamoto, S., Krainc, D., Sherman, K. & Lipton, S. A. Antiapoptotic role of the p38 mitogen-activated protein kinase-myocyte enhancer factor 2 transcription factor pathway during neuronal differentiation. *Proc. Natl Acad. Sci. USA* **97**, 7561–7566 (2000).
- Smith, D. S. & Tsai, L. H. Cdk5 behind the wheel: a role in trafficking and transport? *Trends Cell Biol.* **12**, 28–36 (2002).
- Tu, Y., Tornaletti, S. & Pfeiffer, G. P. DNA repair domains within a human gene: selective repair of sequences near the transcription initiation site. *EMBO J.* **15**, 675–683 (1996).
- Tchou, J. *et al.* Substrate specificity of Fpg protein. Recognition and cleavage of oxidatively damaged DNA. *J. Biol. Chem.* **269**, 15318–15324 (1994).
- Lindahl, T. & Barnes, D. E. Repair of endogenous DNA damage. *Cold Spring Harb. Symp. Quant. Biol.* **65**, 127–133 (2000).
- Bruner, S. D., Norman, D. P., Fromme, J. C. & Verdine, G. L. Structural and mechanistic studies on repair of 8-oxoguanine in mammalian cells. *Cold Spring Harb. Symp. Quant. Biol.* **65**, 103–111 (2000).
- Athas, W. F., Hedayati, M. A., Matanoski, G. M., Farmer, E. R. & Grossman, L. Development and field-test validation of an assay for DNA repair in circulating human lymphocytes. *Cancer Res.* **51**, 5786–5793 (1991).
- Landfield, P. W. & Pitler, T. A. Prolonged Ca²⁺-dependent afterhyperpolarizations in hippocampal neurons of aged rats. *Science* **226**, 1089–1092 (1984).
- Longo, V. D. & Finch, C. E. Evolutionary medicine: from dwarf model to healthy centenarians? *Science* **299**, 1342–1346 (2003).
- Hekimi, S. & Guarente, L. Genetics and specificity of the aging process. *Science* **299**, 1346–1351 (2003).
- Hasty, P., Campisi, J., Hoeijmakers, J., van Steeg, H. & Vijg, J. Aging and genome maintenance: lessons from the mouse? *Science* **299**, 1355–1359 (2003).
- Murphy, C. T. *et al.* Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. *Nature* **424**, 277–283 (2003).
- Ghosh, R. & Mitchell, D. L. Effect of oxidative DNA damage in promoter elements on transcription factor binding. *Nucleic Acids Res.* **27**, 3213–3218 (1999).
- Kyng, K. J., May, A., Kolvaara, S. & Bohr, V. A. Gene expression profiling in Werner syndrome closely resembles that of normal aging. *Proc. Natl Acad. Sci. USA* **100**, 12259–12265 (2003).
- Mecocci, P. *et al.* Oxidative damage to mitochondrial DNA shows marked age dependent increases in human brain. *Ann. Neurol.* **34**, 609–616 (1993).
- McCarroll, S. A. *et al.* Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nature Genet.* **36**, 197–204 (2004).

- Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121 (2001).
- Xu, J. *et al.* Dopamine-dependent neurotoxicity of alpha-synuclein: A mechanism for selective neurodegeneration in Parkinson disease. *Nature Med.* **8**, 600–606 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank G. Verdine and D. Gabuzda for discussions, and M. Yuan, S. Weninger, P. Stieg, P. Dicks, C. Geula and L.-W. Jin for assistance with procurement and dissection of tissue samples. We also acknowledge the Kathleen Price Bryan Brain Bank at Duke University, the Harvard Brain Tissue Resource Center, the Massachusetts General Hospital ADRC Brain Bank, the University of Washington Brain Bank, and the University of Maryland Brain Bank for providing tissue samples. This work was supported in part by grants from the NIH (NIA and NINDS) to B.A.Y. T.L. was supported by a fellowship from the National Institute on Aging, and S.Y.K. was supported by a fellowship from the Leukemia and Lymphoma Society.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.A.Y. (Bruce.Yankner@childrens.harvard.edu).

Mismatch repair genes identified using genetic screens in Blm-deficient embryonic stem cells

Ge Guo, Wei Wang & Allan Bradley

The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

Phenotype-driven recessive genetic screens in diploid organisms require a strategy to render the mutation homozygous. Although homozygous mutant mice can be generated by breeding, a reliable method to make homozygous mutations in cultured cells has not been available, limiting recessive screens in culture. Cultured embryonic stem (ES) cells¹ provide access to all of the genes required to elaborate the fundamental components and physiological systems of a mammalian cell. Here we have exploited the high rate of mitotic recombination in Bloom's syndrome protein (Blm)-deficient ES² cells to generate a genome-wide library of homozygous mutant cells from heterozygous mutations induced with a revertible gene trap³ retrovirus. We have screened this library for cells with defects in DNA mismatch repair (MMR), a system that detects and repairs base–base mismatches⁴. We demonstrate the recovery of cells with homozygous mutations in known and novel MMR genes. We identified *Dnmt1* (ref. 5) as a novel MMR gene and confirmed that *Dnmt1*-deficient ES cells exhibit micro-satellite instability⁶, providing a mechanistic explanation for the role of *Dnmt1* in cancer. The combination of insertional mutagenesis in Blm-deficient ES cells establishes a new approach for phenotype-based recessive genetic screens in ES cells.

In order to establish a genetic background in which homozygous mutations could be readily recovered, we generated mice and ES cells with double-targeted mutations in the *Blm* locus^{2,7}. We have previously established that Blm-deficient cells carrying heterozygous mutations segregate homozygous daughters *in vitro* and *in vivo*, presumably through mitotic recombination between non-sister chromatids^{2,8} (Fig. 1a). The rate of loss of heterozygosity (LOH), determined by Luria–Delbruck fluctuation analysis, is 4.2×10^{-4} events per locus per cell per generation in Blm-deficient ES cells, which is elevated 18-fold compared with wild type cells². In practical terms, a single Blm-deficient ES cell with a heterozygous autosomal mutation will have segregated several

daughter cells with homozygous mutations by the time the cell has expanded to a colony containing 2,000 to 5,000 cells. Thus, by generating and expanding a library of heterozygous mutations in *Blm*-deficient ES through more than 13 population doublings, the library will contain a genome-wide set of homozygous mutations.

We elected to generate a library of mutations in *Blm*-deficient ES cells by using a gene trap vector. Gene trap mutagenesis efficiently generates large numbers of mutations and provides fast molecular access to the mutant allele⁹. To enable use of the β geo reporter¹⁰, we excised the *loxP*-flanked *PGKneo* cassette from double targeted *Blm*(*m1/m3*) ES cells² to generate *Blm*(*m3/m4*) ES cells (Fig. 1b).

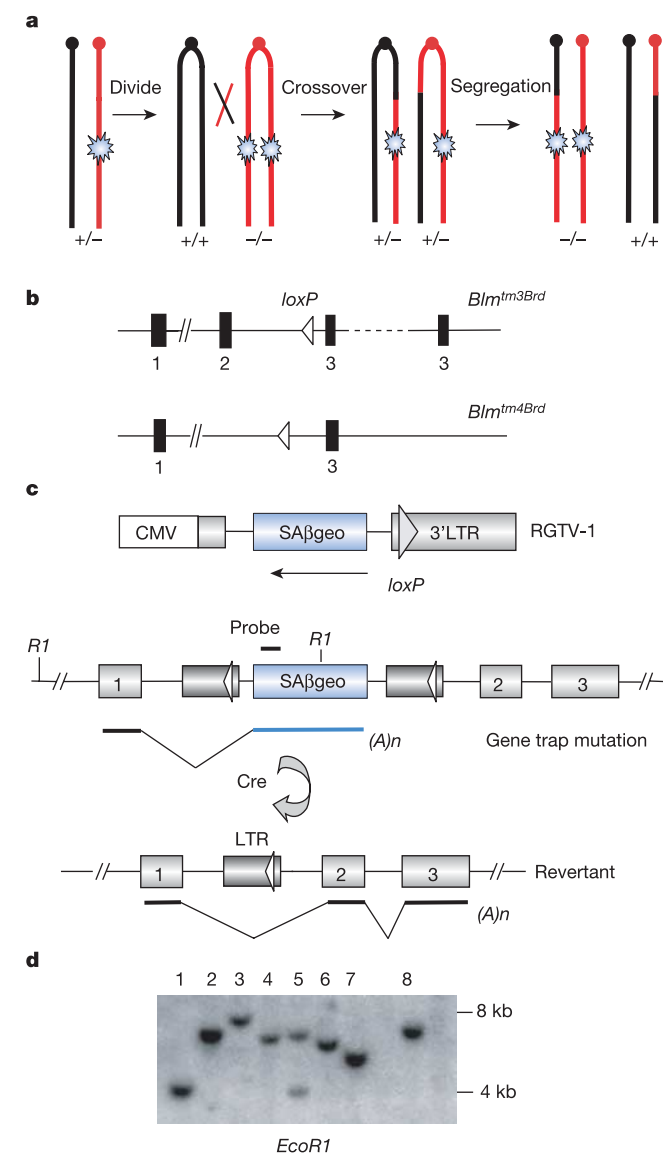


Figure 1 Mechanisms and components used in the screen. **a**, Loss of heterozygosity caused by mitotic recombination. Crossover between non-sister homologous chromatids followed by segregation results in homozygous mutants or wild type cells. **b**, Targeted *Blm* alleles. *Blm*^{tm3Brd}, an allele with exon 3 duplicated, resulting in a frameshift mutation. *Blm*^{tm4Brd}, an allele with exon 2 deleted, derived from *Blm*^{tm1Brd}. **c**, Reversible gene trap mutagenesis. Illustration of the retroviral producer construct, RGTV-1, an integrated provirus in an intron of a gene and the Cre-reverted gene trap locus with a single LTR retained in the locus. R1: *EcoR1*. **d**, Southern blot analysis of gene trap clones demonstrating different host/proviral junction fragment in each clone.

The reversible gene trap retrovirus (RGTV-1, Fig. 1c) was constructed to contain the splice acceptor β geo gene trap cassette in reverse orientation in a retroviral backbone. To facilitate confirmation that a retroviral insertion caused a phenotype, the gene trap insertion is reversible. The integrated provirus carries *loxP* sites in both long terminal repeats (LTRs) and can be excised with Cre, leaving a single LTR in the locus³. As the majority of gene trap insertions are located in introns, excision of the provirus would be expected to reverse the mutant phenotype if the viral insertion was responsible for the phenotype (Fig. 1c).

To demonstrate that recessive screens could be conducted, we performed a screen for MMR mutants. The MMR system is responsible for the repair of mismatched nucleotides arising during DNA replication. The major molecular components of the mismatch recognition machinery are homologues of two *Escherichia coli* genes *MutS* and *MutL* (ref. 11). Although mutation of these genes cause human diseases, such as hereditary non-polyposis colon cancer (HNPCC) (ref. 12), very little is known in eukaryotes about the mechanism of strand discrimination or how the recognition of a mismatch affects cell cycle progression and results in cell death⁴. MMR mutant cell lines are resistant to a number of drugs which modify DNA bases through alkylation (such as *N*-methyl-*N'*-nitro-*N*-nitrosoguanidine) or by incorporation of nucleotide analogues (for example, 6-thioguanine (6-TG)) (ref 13, 14, 15). Notably, 6-TG resistance enables MMR mutants to be directly selected from libraries of mutations.

Selection for MMR mutants using 6-TG requires hypoxanthine-phosphoribosyl-transferase (*Hprt*) activity to convert 6-TG into its toxic phosphorylated nucleotide derivative. The *Blm*-deficient cells were established in AB2.2 ES cells that carry the *hprt*^{mb2} mutation¹⁶. To provide *Hprt* activity in the *Blm*-deficient cells, *Hprt* mini gene cassettes were sequentially targeted into both alleles of the *Gdf9* locus¹⁷ to generate the *Hprt* positive and *Blm*-deficient cell line, *Blm*^{tm3/tm4}, *Gdf9*^{Hprt/Hprt-*loxP*} (NGG)(data not shown).

To establish the mutation library a single cell clone of *Blm*^{tm3/tm4}, *Gdf9*^{Hprt/Hprt-*loxP*} cells (NGG5-3) was expanded and eight pools of NGG5-3 ES cells were infected with RGTV-1 at a multiplicity of less than 0.01 colony forming units per treated cell. The pools were selected in G418 to recover gene trap events with an average of one provirus per selected cell. Each pool was composed of approximately 1,200 independent gene-trap clones with single allele mutations. These pools were expanded for a minimum of 14 population doublings to facilitate the segregation of homozygous mutants. The pools were then selected in 6-TG. A total of 136, 6-TG-resistant clones were recovered from 500 million cells plated. 6-TG-resistant clones from each pool were composed of a variable number of daughter clones from independent mutations represented in the primary pools. The relationship between clones could be established by Southern analysis of the proviral/host junction fragments using a retroviral probe (Fig. 1d).

The gene trap transcript in each clone was recovered using 5'RACE (rapid amplification of cDNA ends) and the proviral/host junction was isolated by a splinkerette polymerase chain reaction (PCR) (ref. 18). We identified 56 independent clones from the 136 clones analysed. Using gene specific flanking probes, the mutated locus was further assessed to determine whether the mutation was homozygous or heterozygous (data not shown). Ten different insertions in three genes were homozygous (Table 1).

The most frequently mutated gene identified in this screen was *Msh6*, a mismatch recognition protein^{19,20}. 5'RACE identified the same junction fragment between β geo and exon 1 of *Msh6* in seven clones (Fig. 2a). The sequence of the proviral/host junction demonstrated the independent origin of each of these clones, as the proviral insertions mapped to seven different positions in intron 1 (Fig. 2b). Southern analysis using an *Msh6* probe (Fig. 2c) confirmed different proviral/*Msh6* junction fragments in each

clone. Notably, none of the clones retained their wild type *Msh6* allele, showing that the most critical feature of this screen was working, namely that all of the insertions were homozygous (Fig. 2c).

The effect of the homozygous proviral insertions on the expression of *Msh6* was assessed by northern blot analysis (Fig. 2d). All of the clones examined expressed only trace levels of *Msh6* (possibly from feeder cells). The reversibility of the mutations in *Msh6* was tested by transient expression of Cre. The reverted clones were all 6-TG-sensitive (Fig. 2e, f) and *Msh6* transcript levels returned to normal (Fig. 2d). Thus, the RGTV-1 virus can effectively mutate the locus, provide molecular access to the gene and is reversible with Cre.

Retroviral insertions were identified in the first intron of *Dnmt1* (DNA (cytosine-5)-methyltransferase 1) (ref. 5). Southern analysis using a *Dnmt1* probe (Fig. 3a, b) confirmed different proviral/*Dnmt1* junction fragments in two mutant clones and in both cases the gene trap insertions were homozygous. *Dnmt1* expression could not be detected in these clones by either northern blot or PCR with reverse transcription (RT-PCR) (Fig. 3c, d). Excision of the proviral insertion with Cre restored *Dnmt1* expression to wild type levels and the cells became sensitive to 6-TG (Fig. 3e).

Mutation of MMR genes leads to increased microsatellite instability (MSI) (ref. 6). A slippage construct (p-Slip) was generated to quantitatively measure MSI (Fig. 4a). p-Slip contains 17 dinucleotide repeats downstream from the initiation codon of the puromycin resistance gene, so that the *Puro* coding sequence is out of frame. Genomic instability, resulting in loss of one or gain of two repeat units, will restore the reading frame and activate the puromycin resistance gene. To accurately compare the MSI rate between clones, a single copy of the p-Slip cassette was targeted into the *ROSA26* locus²¹ in the mutants, *Msh6*-V1 and *Dnmt1*-V1 as well as two control cell lines, AB2.2 and *Blm*-deficient cells (data not shown). Independent targeted clones were identified and single cell clones were expanded and analysed for puromycin resistance. As expected, the *Msh6*-V1 cells exhibited a higher mutation rate than AB2.2 cells. The MSI rate in *Dnmt1*-V1 cells was only marginally higher than wild type, but there was evidence of a genetic interaction between *Dnmt1* and *Blm*. In order to separate the role of *Dnmt1* from *Blm*, we created a *Dnmt1*-deficient cell line, *Dnmt1*-KO, by conventional gene-targeting in *Blm*-wild type cells (Fig. 4b, c). The targeted *Dnmt1*-deficient cells exhibited a rate of MSI that was four times higher than the MSI rate in wild type cells and half of that in *Msh6* mutants (Fig. 4d). Thus, by these criteria, *Dnmt1* seems to qualify as a component of the MMR system.

Dnmt1 has not previously been implicated in MMR, however deficiency of *Dnmt1* causes an increase in the mutation rate²². Mice with *Dnmt1* mutations are tumour susceptible and display chromosome instability^{23,24}. We tested whether the hypomethylated genome in our *Dnmt1*-deficient cells was causing MSI by an indirect change in gene expression of MMR genes. Expression of *Msh2*, *Msh6*, *Msh3*, *Mlh1* and *Pms2*, was not less than wild type levels (data not shown).

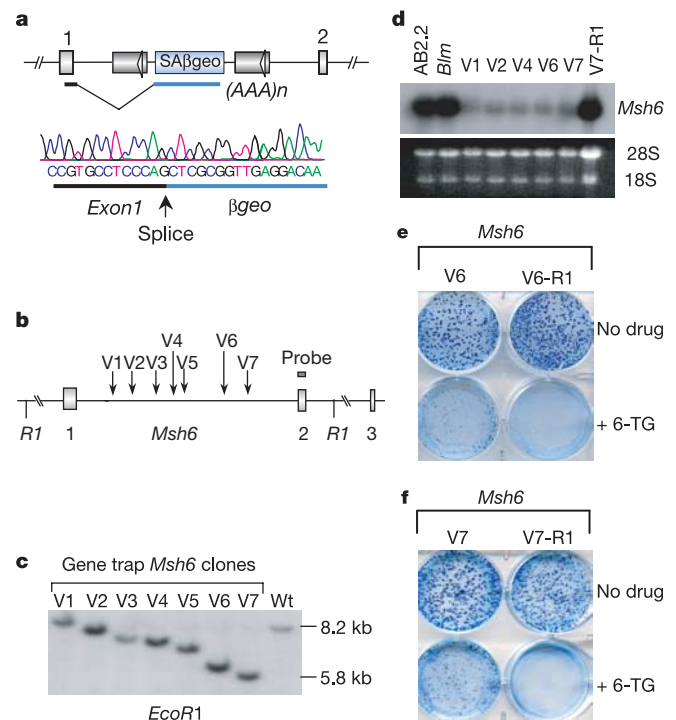


Figure 2 Molecular analysis of *Msh6* gene trap mutants. **a**, Fusion transcript of *Msh6* exon1 with β geo. **b**, Position of retroviral integration sites in intron 1 of *Msh6* clones, V1–V7. R1: *EcoR1*. **c**, Southern blot analysis of proviral/host junctions using an *Msh6* probe. Note that the viral insertions are homozygous, demonstrated by the absence of the wild-type fragment. **d**, Northern blot analysis of *Msh6* expression. AB2.2 cells and *Blm*-deficient ES cells were used as controls. V7-R1 is a Cre-reverted line derived from the mutant *Msh6*-V7. **e**, **f**, Low density survival assay demonstrating 6-TG-resistance in two mutant clones (*Msh6*-V6, *Msh6*-V7) and sensitivity of the revertants (*Msh6*-V6-R1, *Msh6*-V7-R1).

Although our experiments do not define the precise role of *Dnmt1* in mammalian MMR, recognition of hemi-methylated DNA by MutH is central to MMR in *E.coli*¹¹.

The design of the screen was validated by the identification of known and novel components of the MMR pathway. However, the screen seems unsaturated because mutations in known MMR genes (for example *Msh2* and *Mlh1*) were not identified, despite their known resistance to 6-TG. The recovery of seven independent *Msh6* mutants suggests good coverage of the genome, as *Msh6* does not appear to be an integration hot-spot in gene trap databases. The failure to identify mutations in *Msh2* and *Mlh1* suggests that there may be functional interactions, which prevents the isolation of *Msh2* or *Mlh1* mutants in a *Blm*-deficient background. It is expected

Table 1 Gene trap 6-TG resistant clones

Gene	Number of independent clones	Chromosome	Genotype	Revertible	Proviral location
<i>Msh6</i> *	7	Chr17	–/–	Yes	Intron
<i>Dnmt1</i>	2	Chr9	–/–	Yes	Intron
<i>Tgif</i>	1	Chr17	–/–	No	Intron
<i>Sp1</i>	1	Chr15	+/-	NA	5' UTR
<i>ENSMUSG00000032361</i>	1	Chr9	+/-	Yes	Intron
<i>Adprt12</i> *	1	Chr14	ND	ND	Intron
Total	13				

Gene trap clones with different proviral insertion sites in the same gene are grouped. Gene trap lines: homozygous '–/–'; heterozygous '+/-'. One clone with a viral insertion site in the 5' LTR region is designated as 'NA' as the retroviral LTR remains in the 5' UTR region after Cre-mediated recombination and this might disrupt the genes' expression. ND, not determined. Gene name, ID number, chromosome location and intron/exon information were analysed based on the Ensembl database (ver m30, <http://www.ensembl.org>).

*Genes identified in *Caenorhabditis elegans* screen³⁰.

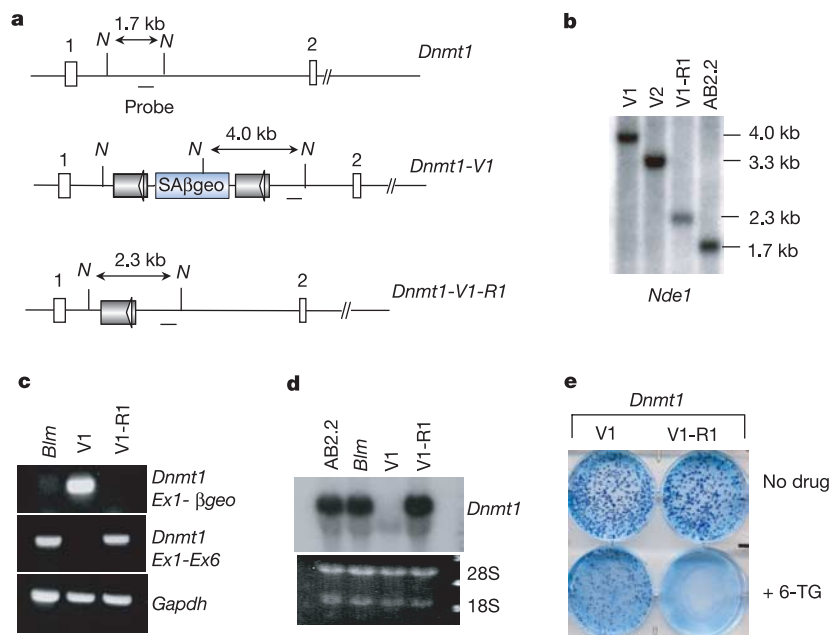


Figure 3 Analysis of *Dnmt1* gene trap mutants. **a**, Schematic of 5' portion of *Dnmt1* in wild type, the *Dnmt1-V1* gene trap and Cre-reverted allele, *Dnmt1-V1-R1*. **b**, Southern blot analysis of the wild type (AB2.2), homozygous mutants *Dnmt1-V1* (V1), *Dnmt1-V2* (V2) and the Cre revertant, *Dnmt1-V1-R1* (V1-R1) cells. **c**, RT-PCR analysis of *Dnmt1*, showing the fusion transcript and the absence of *Dnmt1* expression in the *Dnmt1* gene-

trap mutant and normal expression in the Cre-revertant. **d**, Northern blot analysis of *Dnmt1* expression in *Dnmt1* gene trap mutants and a revertant. **e**, Low density 6-TG survival assay demonstrating 6-TG-resistance (*Dnmt1-V1*) and sensitivity (*Dnmt1-V1-R1*).

that gene trap mutagenesis with a single vector would give incomplete genome coverage. Undoubtedly, RGTV-1 should be considered as effective but limited in its coverage. A larger collection of different insertional mutagens will be required to achieve broader genome coverage.

A portion of the clones recovered were false positives (see supplementary data). These could be identified because the mutations were either heterozygous or non-revertible, or both. False positives can arise from random mutations in MMR genes that have become homozygous. We identified homozygous genomic rearrangement of *Msh6* in three background clones by Southern analysis using cDNA probes. However, no changes in *Msh2* and *Mlh1* were identified (data not shown). In principle, one might anticipate a reduced background by transient knockdown of *Blm*²⁵. However, the greatest contributor to background is likely to be caused by heterozygous mutations that pre-exist in the population of the cells used to generate the library at the time of retroviral infection. Evidently homozygous mutants in MMR genes were not present in these cells, otherwise, the background would have been overwhelming. Thus, it does not seem likely that transient knockdown of *Blm* will reduce this background. The *Blm* mutation used in this study does not have overt genetic consequences on cells or mice². In isogenic ES cells and inbred mice, elevated mitotic recombination between homologous sister chromatids will be genetically neutral.

With the completion of the mouse genome sequence, strategies that enable high throughput genome-wide genetic analysis of gene function are required. In this study, we have developed a strategy to generate libraries of homozygous recessive mutations in cultured ES cells, using *Blm*-deficient cells and a revertible gene trap retrovirus. These libraries enable recessive genetic screens to be conducted. In a parallel study, the chemical mutagen ENU (*N*-ethyl-*N*-nitrosourea) has been used, in combination with a conditional allele of *Blm*, to recover mutations in the glycosylphosphatidylinositol (GPI)-

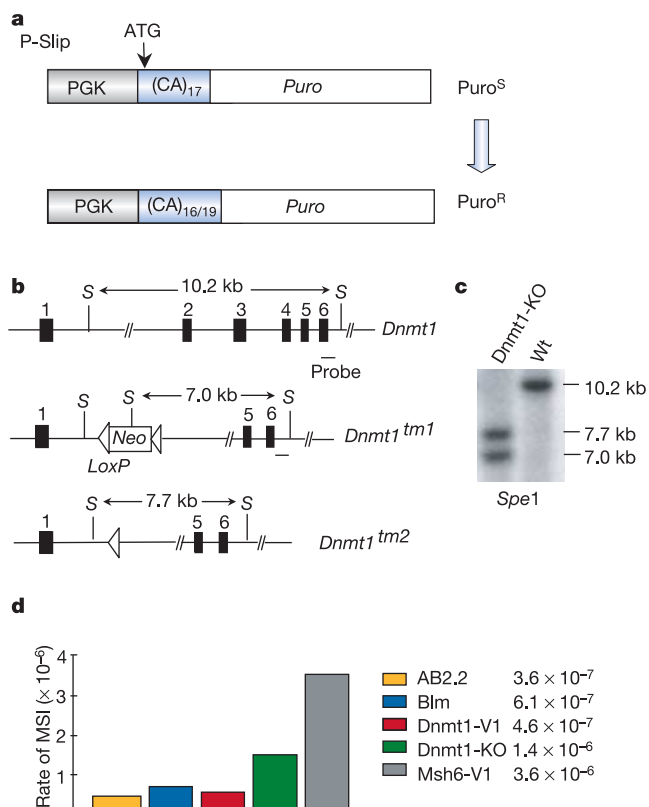


Figure 4 Slippage assay for micro-satellite instability. **a**, Schematic of Puro slippage cassette, p-Slip. Puro^S, puromycin sensitive; Puro^R, puromycin resistant. **b**, Schematic of the targeted disruption of *Dnmt1*. **c**, Southern blot analysis of the gene targeted *Dnmt1*-KO (*tm1/tm2*) ES cells. **d**, Rate of MSI of the targeted p-Slip cassette determined by Luria-Delbruck fluctuation analysis.

anchor biosynthesis pathway²⁵. The use of ENU mutagenesis has the advantage of greater genome coverage but does not allow high throughput identification of new genes. This is because expression cloning needs to be used to identify the gene after the mutant clones have been identified. The development of the reversible insertional gene trap mutagen greatly facilitates the rapid identification and confirmation of the molecular basis of the selected phenotype. □

Methods

Hprt^{+/+}, *Blm*-deficient ES cells

Blm-deficient ES cells were derived from the cell line, *Blm*^{tm1Brd/tm3Brd} (*m1/m3*), described previously². The *loxP*-flanked *PGKneo* selection cassette in *Blm*^{tm1} allele was deleted by Cre-*LoxP* mediated recombination to generate the *Blm*^{tm4Brd} allele. Two autosomal *Hprt* alleles were introduced into *Blm* (*m3/m4*) cells by targeting the *PGKHprt* minigenes into both alleles of the *Gdf9* locus.

Construction of the gene trap vector, RGTV-1

A *loxP* site was inserted into the 3' LTR of the pBabe retroviral vector²⁶, deleting a 267 base pair (bp) *Xba*I/*Nhe*I viral enhancer fragment. The retroviral backbone was PCR amplified and cloned into pcDNA3 (Invitrogen) expression vector to generate the retroviral vector transcribed by the CMV (cytomegalovirus) promoter. The *SAβgeo* gene trap cassette was cloned into the retroviral backbone in reverse transcriptional orientation with respect to the CMV promoter.

Generation of gene trap mutations

A single cell clone of NKG5-3 was expanded. Three million cells were seeded onto 16 90 mm feeder plates on day 0. From day 1, the seeded ES cells were infected with 6 ml of supernatant from retroviral producer cells (with a titer of 30 colony forming units per ml) every 5 h for 2 d. G418 (180 µg ml⁻¹ active ingredient) selection was initiated on day 4 and continued for 6 d until individual ES cell clones were visible. On day 10, the G418 resistant clones within each plate were pooled and expanded for 4 d before initiating 6-TG selection.

Screening for 6-TG resistant ES cells

The plating density used for the screen was optimized in pilot experiments by testing the seeding densities required to recover MMR deficient cells (*Msh2*) while killing *Hprt*-negative mutants as a result of metabolic co-operation with adjacent *Hprt* positive cells²⁷. Five million cells were seeded per 90 mm feeder plate and 2 µM, 2-amino-6-mercaptopurine (6-TG, Sigma) selection was initiated 24 h after plating and continued for 8 d. 6-TG resistant clones were allowed to grow for 4 more days without selection before colonies were picked and expanded for molecular analysis.

Isolation of the proviral junction

Isolation of the proviral junction was performed using the splinkerette-PCR method as described¹⁸. PCR products were gel purified and 5–20 ng was used for each sequencing reaction. Sequencing primers: HMP2-seq (5'-GGTGTCGACACTAGTGG-3') and HM001-seq (5'-ACAGGTGGGGTCTTCA-3').

Rapid isolation of the 5' end of gene trap transcripts by 5' RACE

5' RACE reactions were performed with the 5' RACE system (Invitrogen). Primers for 5' RACE reaction and sequencing: LacZ-GSP1, 5'-GGGCCTCTTCGCTATTACGC-3'; LacZ-GSP2, 5'-ATGTGCTGCAAGGCGATTAAAG-3'; SA-GSP3, 5'-GTTGTAAACGACGGGATCCGCCAT-3'; SA-seq, 5'-TGTCACAGATCATCAAGCTTATC-3'.

Cre reversal

5,000,000 cells were electroporated with 10 µg Cre expressing plasmid and plated at low density on 90 mm feeder plates to allow the growth of individual ES cell colonies. The deletion of the gene trap retrovirus was identified by PCR, using primers that amplify the *lacZ* portion of the βgeo cassette and confirmed by Southern blot using a *LacZ* probe. 6-TG resistance was assessed by low density plating. Two to three hundred ES cells were plated in 30 mm tissue culture plates and allowed to grow for 10 d to form ES cell clones. Each individual clone was visualized by staining with 2% methylene blue in 70% ethanol.

Targeted disruption of *Dnmt1*

A PAC genomic clone containing the mouse *Dnmt1* locus was isolated from the mouse129/SvEv PAC genomic library, RPCI-21 PAC. A replacement targeting vector (*Dnmt1*-TV1) was generated by *E. coli* recombination²⁸ to delete exons 2, 3 and 4 of the *Dnmt1* gene. *Dnmt1*-TV1 was targeted into AB2.2 ES cells to generate the double knockout *Dnmt1*^{tm1/tm2} ES cells using the marker-recycling method.

Construction of the ROSA-p-Slip

Oligonucleotides containing (CA) repeats were synthesized: *Bam*H1-(CA)₁₇, (5'-CTAGTGTATC(TG)₁₇ TTG-3') and *Spe*I-(CA)₁₇, (5'-GATCCAA(CA)₁₇ GATACA-3'). The *PGK* promoter with a translation initiation site ATG and the *PurobpA* coding sequences were separately amplified by PCR from a *PGKPurobpA* cassette. The two fragments and the oligonucleotide were ligated into pBluescript (Stratagene) to create the p-Slip cassette. *PGKEM7/BS*D (a gift from P. Liu) and the p-Slip cassette were ligated into *Nhe*I/*Bgl*III site of pROSA26-1 plasmid to create the Rosa-p-Slip targeting vector.

Fluctuation analysis of the MSI

ES cell lines with the targeted p-Slip cassette were identified by Southern analysis and single cell clones were isolated. 12–24 single cell clones from each line were expanded to the desired number of cells. The expanded clones were trypsinized, counted, and plated in medium with puromycin (2 µg ml⁻¹) and selected for 8 d. Drug resistant colonies were stained with methylene blue and counted. The mutation rate was calculated by the Luria-Delbruck method of means²⁹.

Received 28 January; accepted 13 May 2004; doi:10.1038/nature02653.

- Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154–156 (1981).
- Luo, G. *et al.* Cancer predisposition caused by elevated mitotic recombination in Bloom mice. *Nature Genet.* **26**, 424–429 (2000).
- Ishida, Y. & Leder, P. RET: a poly A-trap retrovirus vector for reversible disruption and expression monitoring of genes in living cells. *Nucleic Acids Res.* **27**, e35 (1999).
- Buermeyer, A. B., Deschenes, S. M., Baker, S. M. & Liskay, R. M. Mammalian DNA mismatch repair. *Annu. Rev. Genet.* **33**, 533–564 (1999).
- Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915–926 (1992).
- Strand, M., Prolla, T. A., Liskay, R. M. & Petes, T. D. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365**, 274–276 (1993) [Erratum in *Nature* (368), 569 (1994)].
- Ellis, N. A. *et al.* The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell* **83**, 655–666 (1995).
- Groden, J., Nakamura, Y. & German, J. Molecular evidence that homologous recombination occurs in proliferating human somatic cells. *Proc. Natl Acad. Sci. USA* **87**, 4315–4319 (1990).
- Stryke, D. *et al.* BayGenomics: a resource of insertional mutations in mouse embryonic stem cells. *Nucleic Acids Res.* **31**, 278–281 (2003).
- Friedrich, G. & Soriano, P. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev.* **55**, 1513–1523 (1991).
- Modrich, P. & Lahue, R. Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annu. Rev. Biochem.* **65**, 101–133 (1996).
- Fishel, R. *et al.* The human mutator gene homolog *MSH2* and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027–1038 (1993).
- Branch, P., Aquilina, G., Bignami, M. & Karran, P. Defective mismatch binding and a mutator phenotype in cells tolerant to DNA damage. *Nature* **362**, 652–654 (1993).
- Swann, P. F. *et al.* Role of postreplicative DNA mismatch repair in the cytotoxic action of thioguanine. *Science* **273**, 1109–1111 (1996).
- Abuin, A., Zhang, H. & Bradley, A. Genetic analysis of mouse embryonic stem cells bearing *Msh3* and *Msh2* single and compound mutations. *Mol. Cell. Biol.* **20**, 149–157 (2000).
- Kuehn, M. R., Bradley, A., Robertson, E. J. & Evans, M. J. A potential animal model for Lesch-Nyhan syndrome through introduction of HPRT mutations into mice. *Nature* **326**, 295–298 (1987).
- Dong, J. *et al.* Growth differentiation factor-9 is required during early ovarian folliculogenesis. *Nature* **383**, 531–535 (1996).
- Mikkers, H. High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. *Nature Genet.* **32**, 153–159 (2002) [Erratum in *Nature Genet.* **32**, 459 (2002)].
- Palombo, F. *et al.* GTBP, a 160-kilodalton protein essential for mismatch binding activity in human cells. *Science* **268**, 1912–1914 (1995).
- Papadopoulos, N. *et al.* Mutations of GTBP in genetically unstable cells. *Science* **268**, 1915–1917 (1995).
- Zambrowicz, B. P. *et al.* Disruption of overlapping transcripts in the ROSA beta geo 26 gene trap strain leads to widespread expression of beta-galactosidase in mouse embryos and hematopoietic cells. *Proc. Natl Acad. Sci. USA* **94**, 3789–3794 (1997).
- Chen, R. Z. *et al.* DNA hypomethylation leads to elevated mutation rates. *Nature* **395**, 89–93 (1998).
- Gaudet, F. *et al.* Induction of tumors in mice by genomic hypomethylation. *Science* **300**, 489–492 (2003).
- Eden, A., Gaudet, F., Waghmare, A. & Jaenisch, R. Chromosomal instability and tumors promoted by DNA hypomethylation. *Science* **300**, 455 (2003).
- Yusa, K. *et al.* Genome-wide phenotype analysis in ES cells by regulated disruption of Bloom's syndrome gene. *Nature* (this issue).
- Morgenstern, J. & Land, H. Advanced mammalian gene transfer: high titre retroviral vectors with multiple drug selection markers and a complementary helper-free packaging cell line. *Nucleic Acids Res.* **18**, 3587–3596 (1990).
- Hooper, M. L. & Slack, C. Metabolic co-operation in HGPRT⁺ and HGPRT⁻ embryonal carcinoma cells. *Dev. Biol.* **55**, 271–284 (1977).
- Liu, P., Jenkins, N. A. & Copeland, N. G. A highly efficient recombineering-based method for generating conditional knockout mutations. *Genome Res.* **13**, 476–484 (2003).
- Luria, S. E. & Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491–511 (1943).
- Pothof, J. *et al.* Identification of genes that protect the *C. elegans* genome against mutations by genome-wide RNAi. *Genes Dev.* **17**, 443–448 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We would like to thank G. Luo and X. Wang for discussions at an early phase in the design of this screen, F. Law and A. Beasley for help with tissue culture and B. Skarnes, P. Liu, D. Adams and J. Jonkers for their comments on this manuscript. This work was supported by the Wellcome Trust.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.B. (abradley@sanger.ac.uk).

Genome-wide phenotype analysis in ES cells by regulated disruption of Bloom's syndrome gene

Kosuke Yusa¹, Kyoji Horie^{1,2}, Gen Kondoh¹, Michiyoshi Kouno¹, Yusuke Maeda³, Taroh Kinoshita³ & Junji Takeda^{1,2}

¹Department of Social and Environmental Medicine, Graduate School of Medicine, ²Collaborative Research Center for Advanced Science and Technology, and ³Department of Immunoregulation, Research Institute for Microbial Diseases, Osaka University, Suita, Osaka 565-0871, Japan

The chief limitation of phenotype-based genetic screening in mammalian systems is the diploid nature of the genome. Cells deficient in the Bloom's syndrome gene (*Blm*) show an increased rate of loss of heterozygosity^{1–3}. Here we have used a tetracycline-regulated *Blm* allele (*Blm*^{tet}) to introduce bi-allelic mutations across the genome in mouse embryonic stem (ES) cells. Transient loss of *Blm* expression induces homologous recombination not only between sister chromatids but also between homologous chromosomes. We considered that the phenotype of ES cells bearing bi-allelic mutations would be maintained after withdrawal of the tetracycline analogue doxycycline. Indeed, a combination of *N*-ethyl-*N*-nitrosourea mutagenesis and transient loss of *Blm* expression enabled us to generate an ES cell library with genome-wide bi-allelic mutations. The library was evaluated by screening for mutants of glycosylphosphatidylinositol-

anchor biosynthesis, which involves at least 23 genes distributed throughout the genome. Mutants derived from 12 different genes were obtained and two unknown mutants were simultaneously isolated. Our results indicate that phenotype-based genetic screening with *Blm*^{tet} is very efficient and raises possibilities for identifying gene functions in ES cells.

Because multipotential ES cells can differentiate into various organs or tissues, 'knock-out' mice can be generated from ES cells by means of gene targeting. *In vitro* differentiation of ES cells to many different cell types such as haematopoietic cells, neurons and cardiomyocytes^{4–6} has been reported, suggesting the possibility of therapeutic applications^{7,8}. The generation of an ES cell library bearing mutations in both alleles (bi-allelic mutations) should have a major impact on analyses of the molecular mechanism of differentiation and the pluripotency of ES cells.

Figure 1a shows the *Blm* allele under conditional regulation of tetracycline (*Blm*^{tet}). Tetracycline-system-based regulatory cassettes (tet cassettes)⁹ were inserted immediately upstream of the translational initiation codons of both alleles of *Blm* to change them into *Blm*^{tet}. Targeting was confirmed by Southern blot analysis (Fig. 1b). We considered that continuous deficiency of *Blm* would cause continuous accumulation of bi-allelic mutations in the genome, resulting in a change in phenotype during long-term culture, while transient loss of *Blm* caused by *Blm*^{tet} would minimize changes in the phenotype.

Regulation of *Blm* expression was examined by using the tetracycline analogue, doxycycline (dox). Addition of dox resulted in a rapid reduction in *Blm* protein (Fig. 1c, d). Notably, the *Blm* protein regained its original expression after the withdrawal of dox (Fig. 1c).

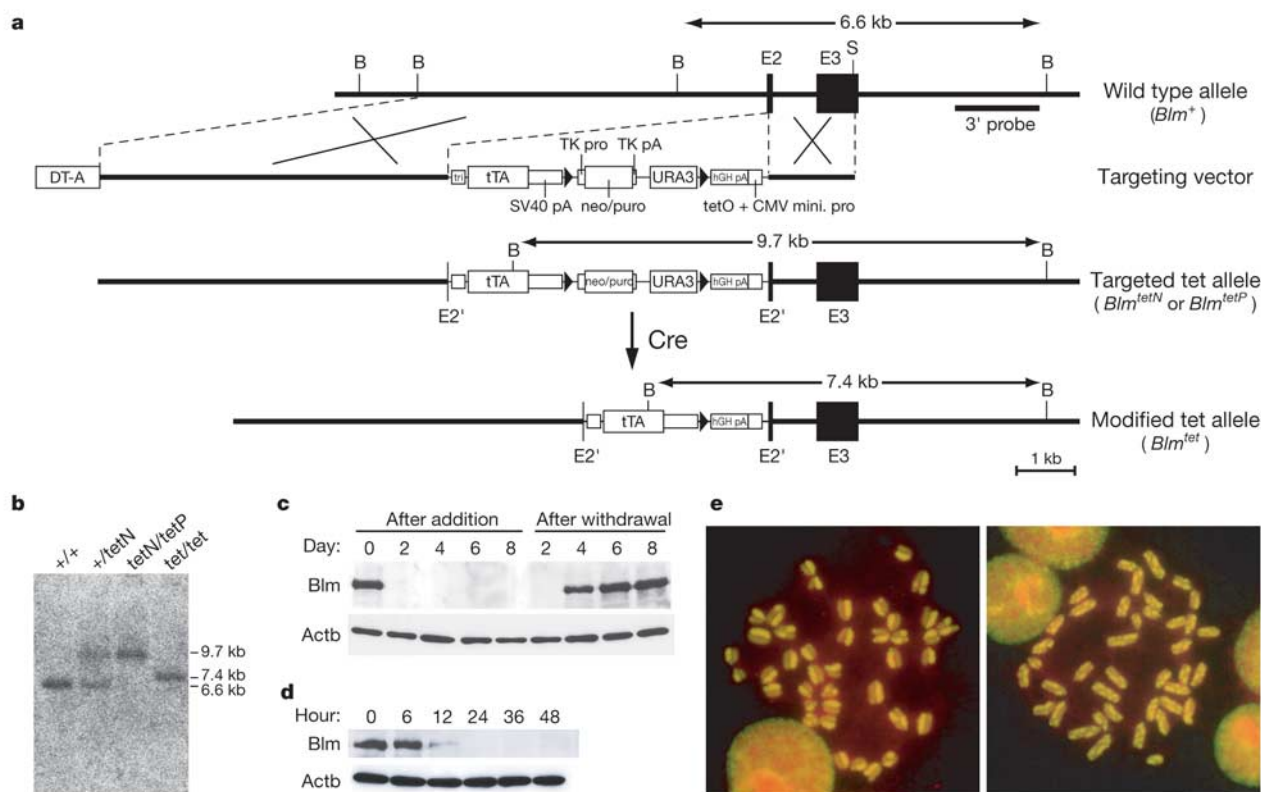


Figure 1 Generation of conditional *Blm* alleles in ES cells and elevation of SCE.

a, Targeting strategy and resulting *Blm* alleles. A tet cassette containing the *neo* or *puro* gene was inserted upstream of the translational initiation site of *Blm* to generate *Blm*^{tetN} or *Blm*^{tetP}, respectively. After Cre expression, these selection markers were deleted, resulting in *Blm*^{tet}. B, *Bam*HI; S, *Sac*I. Abbreviations for the tet cassette are described in

ref. 9. **b**, Southern blot analysis of targeted clones. Genomic DNA was digested with *Bam*HI, separated by electrophoresis and hybridized with the radiolabelled probe shown in **a**. **c**, **d**, Long-term (**c**) and short-term (**d**) analysis of *Blm* expression in *Blm*^{tet/tet} ES cells by western blot. Expression of β -actin (Actb) was used as a loading control. **e**, SCE of dox-treated (right) or non-treated (left) *Blm*^{tet/tet} ES cells.

Increased numbers of sister chromatid exchanges (SCEs), a typical cytogenetic phenomenon of Bloom's syndrome cells¹, were observed (Fig. 1e), while Blm proteins were undetectable (Fig. 1c).

SCE is closely coupled to homologous recombination in vertebrate cells¹⁰; therefore, dox treatment is expected to induce recombination between homologous chromosomes and, when it occurs at the 4n stage, a mono-allelic mutation will become bi-allelic after cell division (Fig. 2a). To examine the effect of transient Blm deficiency on the rate of bi-allelic mutation, we introduced, by means of gene targeting, a mutant *neo* gene into the *Fas ligand* (*FasL*) locus as a model of 'mono-allelic mutation'. We have previously shown that duplication of the mutant *neo* gene at this locus, which represents 'bi-allelic mutation', can be selected by high doses of G418 (ref. 11). The rate of the duplication determined by Luria–Delbruck fluctuation analysis¹² was 8.5×10^{-6} events per cell per generation in the absence of dox, but increased to 2.3×10^{-4} events per cell per generation in the presence of dox (Fig. 2b). Therefore, transient loss of Blm causes a 27-fold increase in the rate of bi-allelic mutation (Fig. 2b). The rates obtained in this study are similar to those reported previously for loss of heterozygosity (LOH) in wild-type (2.3×10^{-5}) and *Blm*-deficient (4.2×10^{-4}) ES cells³.

To determine the chromosomal locations of the crossover, we used polymorphic markers in chromosome 1 of the R1-ES cell line (Fig. 2c), which was established from an F₁ embryo obtained from the breeding of two different inbred 129 substrains (129X1/SvJ $\times$ 129S1/SvImJ)¹³. The respective locations of the *D1Mit1001* and *D1Mit292* markers are located roughly 30 megabases (Mb) proximal and distal from the *FasL* locus. Twenty-eight high-dose G418-resistant clones were chosen to determine the chromosomal locations of the crossover, and ten (~35%) crossovers were found to

occur in a region 30-Mb proximal from the *FasL* locus (Fig. 2c).

The observation of increased mitotic recombination in *Blm*-modified ES cells prompted us to examine the possibility of establishing an ES cell library containing the bi-allelic mutations throughout the genome. *N*-Ethyl-*N*-nitrosourea (ENU) was used as a mutagen because of its extremely high mutagenicity in ES cells¹⁴. To determine whether the distribution and complexity of bi-allelic mutations were high enough to cover the whole genome, we screened for mutant ES cells deficient in glycosylphosphatidylinositol (GPI)-anchor biosynthesis. At least 23 genes, which are widely distributed in the mouse genome, are involved in this pathway. Because mutations in any gene in the GPI pathway yield a deficiency of GPI-anchored proteins on the cell surface, cells deficient in the GPI-anchor can be positively selected by using aerolysin, which kills cells with GPI-anchors¹⁵.

The ES cell line used in this study was of male origin, whereas the *PigA* gene involved in the first step of GPI-anchor biosynthesis is localized on the X chromosome. Therefore, functional disruption of *PigA* does not require bi-allelic mutation and most of the GPI-anchor-deficient mutants would originate from *PigA* mutation¹⁶. To avoid such a bias, extra copies of the *PigA* complementary DNA were introduced into the *Blm*^{tet/tet} ES cells before ENU mutagenesis.

Figure 3a summarizes the protocol for ENU mutagenesis of ES cells, generation of the ES cell library, and screening for GPI-anchor-deficient mutants. We treated 2×10^8 ES cells using an ENU dose of 0.2 mg ml^{-1} for 2 h at 37 °C. Cell viability was roughly 3%, resulting in 6×10^6 cells surviving after ENU treatment. The mutation frequency in surviving cells was 1 in 2,400 at the X-linked mono-allelic hypoxanthine phosphoribosyl transferase (*Hprt*) locus, as determined by selection with 6-thioguanine. It has been reported that ES cells treated at a higher concentration of ENU (0.35 mg ml^{-1}) retain germline competency¹⁴, suggesting that the long-term viability of treated cells resembles that of wild-type ES cells. The frequency of bi-allelic mutations induced by dox treatment for three generations of cell cycles was calculated to be 0.7×10^{-6} per locus, and the number of independent clones

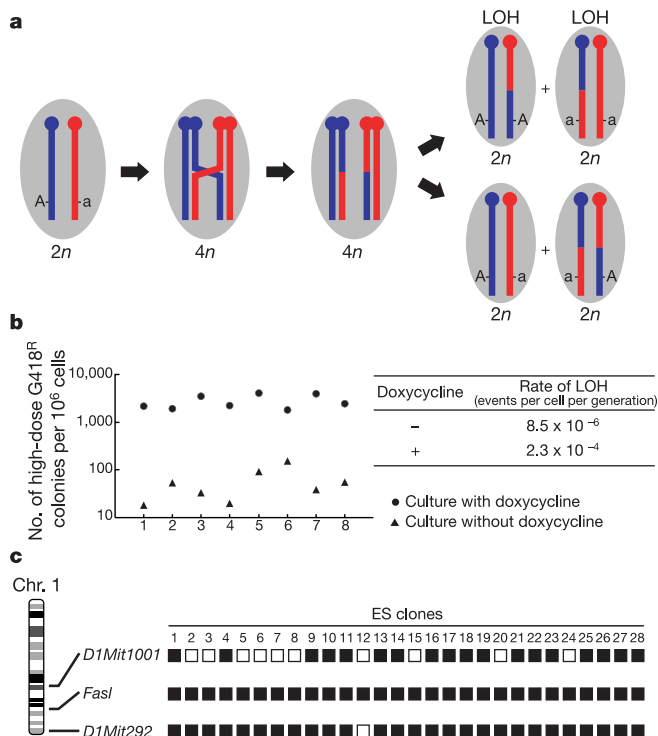


Figure 2 High frequency of LOH in *Blm*-deficient ES cells. **a**, Mechanism of LOH. Heterozygosity is represented as A/a. When homologous recombination occurs at the 4n stage between homologous chromosomes, cells bearing LOH (A/A or a/a) appear after cell divisions. **b**, Luria–Delbruck fluctuation analysis of LOH. Frequency of duplication of the *FasL* locus containing the *neo* gene was examined. **c**, Polymorphic simple sequence length polymorphism (SSLP) marker analysis of 28 bi-allelic mutants of the *FasL* locus. Open and filled squares indicate heterozygosity and LOH, respectively.

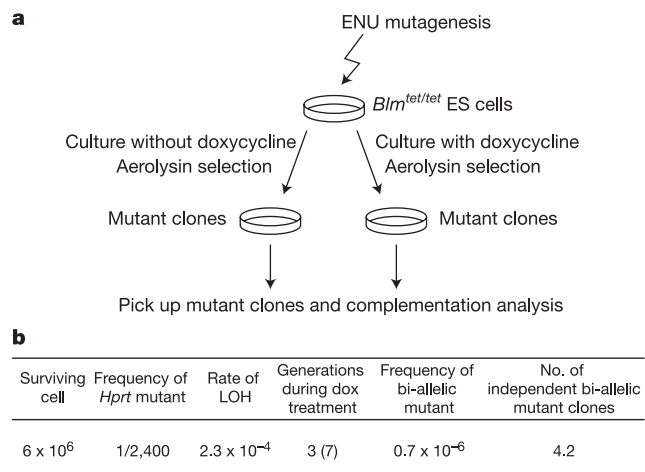


Figure 3 Construction of the mutant ES cell library and screening strategy of GPI-anchor-defective mutants. **a**, Screening of GPI-anchor-defective mutants from the ES library bearing bi-allelic mutations. **b**, Theoretical prediction of effectiveness of this screening strategy. From left to right, the measures are the number of surviving cells after ENU treatment of 2×10^8 ES cells, the frequency of X-linked *Hprt* negative cells measured as a ratio of 6-TG resistant colonies, the mutation rate of LOH as described in Fig. 2b, the number of generations (cell cycles) during dox treatment (number in parentheses shows the total number of cell divisions during dox treatment), the frequency of clones bearing bi-allelic mutation per locus ($2.3 \times 10^{-4} \times 1/2,400 \times 7$), and the number of independent clones bearing bi-allelic mutation per locus after dox treatment.

bearing bi-allelic mutations was estimated at 4.2 per locus (Fig. 3b). These results indicate that our ES cell library contains bi-allelic mutations in most loci.

To verify the practical utility of this principle, we screened the ES cell library with aerolysin, which resulted in the isolation of 35 GPI-anchor-deficient mutants (Fig. 3a). When the GPI-anchored green fluorescent protein (GFP) construct (GFP-GPI)¹⁷ was transfected into wild-type ES cells, GFP-GPI proteins were expressed on the cell surface (Fig. 4a). By contrast, GFP-GPI proteins were expressed on the mutants only when complementary cDNA was supplied (Fig. 4a). These mutants were therefore classified by means of complementation analysis using the transfection of cDNAs of genes involved in GPI-anchor biosynthesis. The mutated genes were found to be distributed widely throughout the genome, and mutants were identified in more than half of the known GPI-anchor biosynthesis genes (12 out of 23) after one round of screening (Fig. 4b).

Figure 4c shows that one mutant was obtained in four genes, but that more than one mutant was obtained in other autosomal genes. Because the genes that were affected frequently had same mutation,

their mutants were probably derived from single clones. Therefore, the difference in mutant numbers would be explained by the differences in the stages at which the mutants were generated after the addition of dox. Sequence analysis of these mutants showed that they did not contain sequences of the wild-type allele (Fig. 4d, e), indicating that bi-allelic mutations had occurred in the ES cells treated with dox. Without dox treatment, only one mutant was isolated (data not shown). In addition, GPI-anchor biosynthesis was not complemented by cDNA transfection in two mutants, which suggested that these mutants had mutations in novel genes involved in the GPI pathway (Fig. 4f). Different genes were mutated in these mutants because the deficiency in GPI-anchor biosynthesis was complemented by cell fusion (data not shown).

We have shown that it is possible to isolate recessive mutants across the genome in mammalian cells. Although we¹¹ and others¹⁸ have reported a method to introduce bi-allelic mutation by Cre-loxP-mediated recombination between homologous chromosomes, that method can be applied only to a pre-selected chromosome that carries loxP sites on both alleles. Many, but not all, loci are functionally haploid in the Chinese hamster ovary (CHO) cell

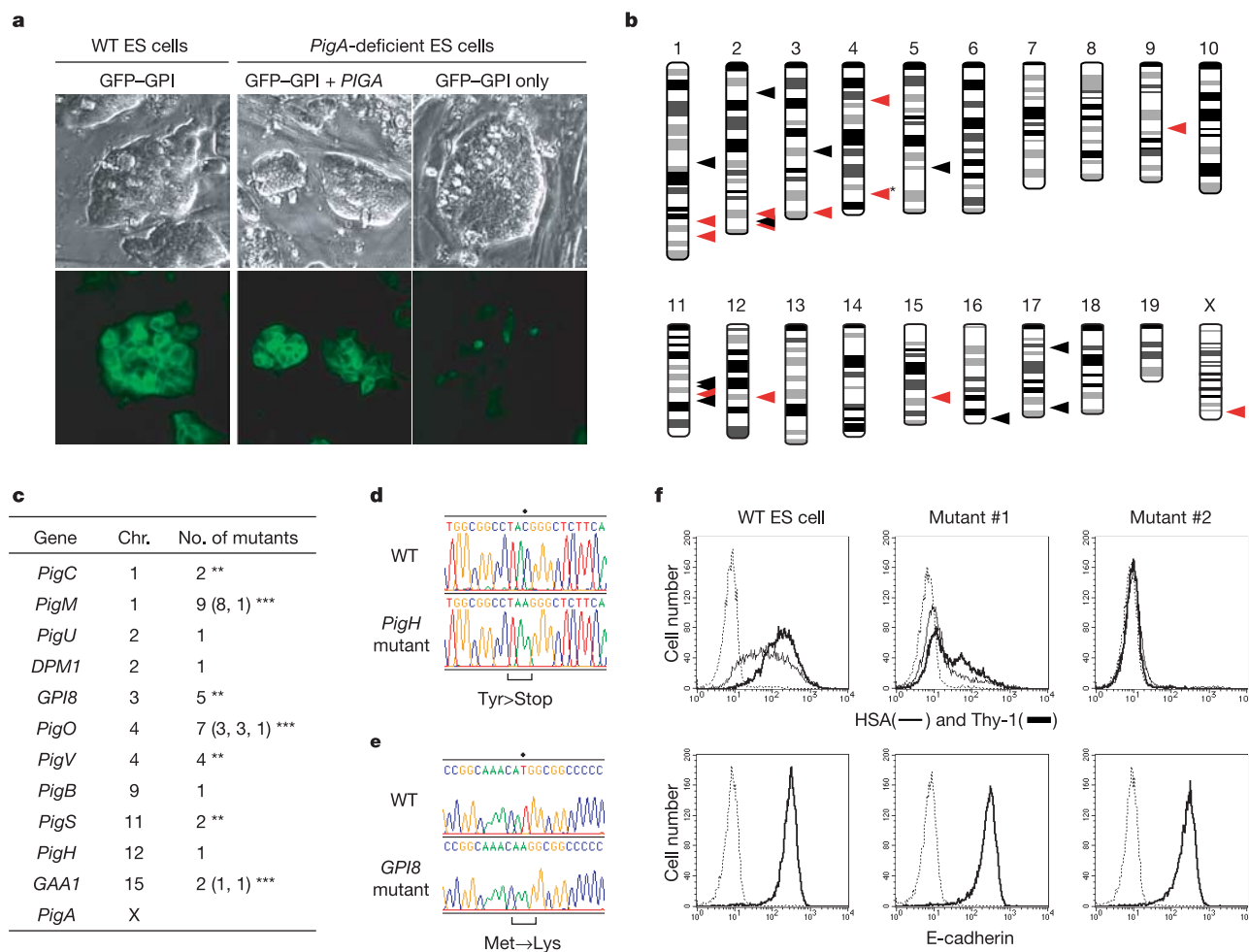


Figure 4 Analysis of GPI-anchor-defective mutants. **a**, Complementation analysis of GPI-anchor-defective mutants. Right, GPI-anchored GFP proteins were expressed on the cell surface of *PigA*-deficient ES cells only when *PigA* cDNA was supplied. Left, wild-type ES cells were transfected with GFP-GPI as a positive control. **b**, Chromosomal locations of 23 genes involved in GPI-anchor biosynthesis. Red arrowheads indicate 12 genes of which mutants were obtained in this screening; black arrowheads indicate 11 genes of which mutants were not obtained; asterisk indicates a cDNA that has been cloned but not published. **c**, Chromosome number of mutated genes and numbers of mutants obtained. The order of mutated genes in each chromosome is given from centromere to telomere.

Mutants contained same mutation; *mutants contained different mutations. Numbers in parentheses indicate number of mutants bearing the same mutation. **d, e**, Sequence analysis of mutations in *PigH* (**d**) and *GPI8* (**e**). Homozygous mutations were verified. **f**, FACS analysis of the two novel GPI mutants. Top, ES cells were stained with biotinylated anti-HSA (thin line) and anti-Thy-1 (thick line) antibodies, followed by streptavidin-phycoerythrin. HSA and Thy-1 are GPI-anchored proteins. Broken lines indicate control staining profiles without biotinylated antibodies. Bottom, expression pattern of a non-GPI-anchored protein, E-cadherin, in the mutants.

line¹⁹, and the isolation of mutant cells with recessive phenotype has been reported²⁰. Because the probability of isolating mutants depends largely on whether target genes are functionally haploid or diploid in CHO cells, non-random isolation of mutants can be expected²¹. In our study, more than half of the genes known to be involved in GPI-anchor biosynthesis could be identified in a single round of selection in ES cells with normal karyotype, thereby verifying the random nature of our selection scheme. Theoretically, bi-allelic mutations occur in most loci (Fig. 3b), but nearly half of the genes involved in GPI-anchor biosynthesis could not be identified. This incomplete coverage may be explained by AT base-pair predominant mutations with ENU in ES cells¹⁴. This possibility can be tested by other chemical mutagens such as EMS and ICR191, which have a different mutation spectrum²².

To identify genes responsible for a given phenotype, expression cloning can be applied. The development of highly efficient systems for cDNA library transduction and recovery, such as an episomal vector stably maintained in ES cells²³ and high-titre retroviral vectors resistant to promoter silencing in ES cells²⁴, will greatly aid the identification of mutated genes. A survey of homologous regions by means of polymorphic markers is likely to narrow down the location of the mutation, as exemplified by Fig. 2c. To achieve fine mapping of mutations, we introduced *Blm*^{tet} alleles into C57BL/6 × 129S4/SvJae F₁ hybrid ES cells²⁵, for which a large number of polymorphic markers are available. Alternatively, tagged mutagenesis such as gene trap²⁶ may be used in place of ENU to facilitate identification of causative genes. In fact, a retroviral gene-trap vector has been used successfully in the accompanying paper for a genome-wide recessive screen²⁷.

For phenotype-based genetic screening, it is essential to establish definitive criteria for judging whether an observed phenotype is caused by genetic mutations or by simply a change in characteristics of wild-type cells. This is especially important in the analysis of ES cells, because a small fraction of an ES cell population may differentiate spontaneously even when culture conditions are carefully controlled. Conditional regulation of *Blm* expression is useful in assessing a given phenotype. If clones with the desired phenotype can be obtained more efficiently under dox-treated than under non-selective culture conditions, then those clones isolated by dox-induced LOH probably contain bi-allelic mutations. Here, the fact that 35 aerolysin-resistant clones were obtained from dox-treated culture, as compared with only 1 clone with non-selective culture, prompted us to characterize those clones further. Because pluripotent ES cells can differentiate into any type of tissue, a method for the comprehensive isolation of bi-allelic mutants should have a major impact on the analysis of molecular mechanism of differentiation *in vitro* as well as *in vivo*. □

Methods

Generation of conditional *Blm* ES cells

Genomic DNA containing the mouse *Blm* gene was isolated from the R1-ES genomic library. The targeting vector was introduced into R1-ES cells and selected by using G418 and/or puromycin. Targeted clones were screened by polymerase chain reaction and confirmed by Southern blot analysis. All targeted clones used in this study possessed a normal karyotype.

Western blot analysis

We cultured *Blm*^{tet/tet} ES cells with 1.0 µg ml⁻¹ of dox and collected them at appropriate time points. To examine *Blm* expression after dox withdrawal, cells cultured in dox-containing media were washed once with PBS and further cultured in the absence of dox until collection. *Blm* protein was detected with an ab476 antibody against BLM (Abcam).

SCE analysis

Blm^{tet/tet} ES cells cultured in the presence or absence of dox for appropriate periods were labelled with 3 µg ml⁻¹ of BrdU for 20 h and treated with 0.1 µg ml⁻¹ of colcemid for 45 min. Chromosome spreads were stained with 0.1 µg ml⁻¹ of acridine orange.

Fluctuation analysis

We introduced the targeting vector for the *FasI* locus with the mutant *neo* gene¹¹ into *Blm*^{tet/tet} ES cells. The rate of bi-allelic mutagenesis was measured by means of Luria–

Delbruck fluctuation analysis as described¹¹. In brief, *Blm*^{tet/tet} ES cells were cultured for 24 h with or without 1.0 µg ml⁻¹ of dox and plated on a 100-mm dish with a clonal-density culture with or without dox to obtain single-cell clones. Expanded single-cell clones were then selected by using high-dose G418 (1.0 mg ml⁻¹) without dox. The number of high-dose G418 resistant clones was counted 10 d after selection.

ENU mutagenesis and screening for GPI-anchor mutants

We carried out ENU mutagenesis and calculated the mutation frequency of the *Hprt* locus as described¹⁴. The mutagenized ES cells were cultured with or without dox for 4 d to generate the mutant ES cell library. The mutant and control libraries were treated in 10 nM proaerolysin (Protox Biotech) in suspension (1.0 × 10⁶ cells per ml) and plated on a gelatin-coated 100-mm dish at 5–8 × 10⁶ cells per dish. The next day, dead cells were washed out and living cells were treated in 5 nM proaerolysin for 8 h before mitomycin-C-treated feeder cells were added to generate GPI-anchor-defective mutant colonies. The resulting colonies were collected, expanded and transfected with GFP–GPI expression vector and cDNAs involved in GPI-anchor biosynthesis for the complementation assay.

Received 2 February; accepted 6 May 2004; doi:10.1038/nature02646.

- German, J. Bloom's syndrome. *Dermatol. Clin.* **13**, 7–18 (1995).
- Groden, J., Nakamura, Y. & German, J. Molecular evidence that homologous recombination occurs in proliferating human somatic cells. *Proc. Natl Acad. Sci. USA* **87**, 4315–4319 (1990).
- Luo, G. *et al.* Cancer predisposition caused by elevated mitotic recombination in Bloom mice. *Nature Genet.* **26**, 424–429 (2000).
- Kyba, M. & Daley, G. Q. Hematopoiesis from embryonic stem cells: lessons from and for ontogeny. *Exp. Hematol.* **31**, 994–1006 (2003).
- Kim, J. H. *et al.* Dopamine neurons derived from embryonic stem cells function in an animal model of Parkinson's disease. *Nature* **418**, 50–56 (2002).
- Parisi, S. *et al.* Nodal-dependent Cripto signaling promotes cardiomyogenesis and redirects the neural fate of embryonic stem cells. *J. Cell Biol.* **163**, 303–314 (2003).
- Reubinoff, B. E., Pera, M. F., Fong, C. Y., Trounstein, A. & Bongso, A. Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. *Nature Biotechnol.* **18**, 399–404 (2000).
- Thomson, J. A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145–1147 (1998).
- Bond, C. T. *et al.* Respiration and parturition affected by conditional overexpression of the Ca²⁺-activated K⁺ channel subunit, SK3. *Science* **289**, 1942–1946 (2000).
- Sonoda, E. *et al.* Sister chromatid exchanges are mediated by homologous recombination in vertebrate cells. *Mol. Cell. Biol.* **19**, 5166–5169 (1999).
- Koike, H. *et al.* Efficient allelic mutagenesis with Cre/loxP-mediated inter-chromosomal recombination. *EMBO Rep.* **3**, 433–437 (2002).
- Luria, S. E. & Delbruck, M. Mutation of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491–510 (1943).
- Lefebvre, L., Dionne, N., Karasikova, J., Squire, J. A. & Nagy, A. Selection for transgene homozygosity in embryonic stem cells results in extensive loss of heterozygosity. *Nature Genet.* **27**, 257–258 (2001).
- Chen, Y. *et al.* Genotype-based screen for ENU-induced mutations in mouse embryonic stem cells. *Nature Genet.* **24**, 314–317 (2000).
- Hong, Y. *et al.* Requirement of N-glycan on GPI-anchored proteins for efficient binding of aerolysin but not Clostridium septicum α-toxin. *EMBO J.* **21**, 5047–5056 (2002).
- Kawagoe, K., Takeda, J., Endo, Y. & Kinoshita, T. Molecular cloning of murine *pig-a*, a gene for GPI-anchor biosynthesis, and demonstration of interspecies conservation of its structure, function, and genetic locus. *Genomics* **23**, 566–574 (1994).
- Kondoh, G. *et al.* Tissue-inherent fate of GPI revealed by GPI-anchored GFP transgenesis. *FEBS Lett.* **458**, 299–303 (1999).
- Liu, P., Jenkins, N. A. & Copeland, N. G. Efficient Cre-loxP-induced mitotic recombination in mouse embryonic stem cells. *Nature Genet.* **30**, 66–72 (2002).
- Gupta, R. S., Chan, D. Y. & Siminovich, L. Evidence for functional hemizygosity at the *Emtr* locus in CHO cells through segregation analysis. *Cell* **14**, 1007–1013 (1978).
- Hanada, K. *et al.* Molecular machinery for non-vesicular trafficking of ceramide. *Nature* **426**, 803–809 (2003).
- Nakamura, N. *et al.* Expression cloning of PIG-L, a candidate N-acetylglucosaminyl-phosphatidylinositol deacetylase. *J. Biol. Chem.* **272**, 15834–15840 (1997).
- Munroe, R. J. *et al.* Mouse mutants from chemically mutagenized embryonic stem cells. *Nature Genet.* **24**, 318–321 (2000).
- Chambers, I. *et al.* Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* **113**, 643–655 (2003).
- Kitamura, T. *et al.* Retrovirus-mediated gene transfer and expression cloning: powerful tools in functional genomics. *Exp. Hematol.* **31**, 1007–1014 (2003).
- Eggan, K. *et al.* Hybrid vigor, fetal overgrowth, and viability of mice derived by nuclear cloning and tetraploid embryo complementation. *Proc. Natl Acad. Sci. USA* **98**, 6209–6214 (2001).
- Friedrich, G. & Soriano, P. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev.* **5**, 1513–1523 (1991).
- Guo, G., Wang, W. & Bradley, A. Mismatch repair genes identified using genetic screens in *Blm*-deficient embryonic stem cells. *Nature* (this issue).

Acknowledgements We thank J. P. Adelman for the tetracycline-system-based regulatory cassette, and V. W. Keng for critically reading the manuscript. This work was supported, in part, by a grant from New Energy and Industrial Technology Development Organization (NEDO) of Japan.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.T. (takeda@mr-envi.med.osaka-u.ac.jp).

Essential role for *de novo* DNA methyltransferase Dnmt3a in paternal and maternal imprinting

Masahiro Kaneda^{1,2}, Masaki Okano^{3,4}, Kenichiro Hata¹, Takashi Sado^{1,5,6}, Naomi Tsujimoto^{1,2,4*}, En Li^{4*} & Hiroyuki Sasaki^{1,2}

¹Division of Human Genetics, Department of Integrated Genetics, National Institute of Genetics, Research Organization of Information and Systems, Mishima 411-8540, Japan

²Department of Genetics, School of Life Science, The Graduate University for Advanced Studies (Sokendai), Mishima 411-8540, Japan

³Laboratory for Mammalian Epigenetic Studies, Center for Developmental Biology, RIKEN, Kobe 650-0047, Japan

⁴Cardiovascular Research Center, Massachusetts General Hospital, Department of Medicine, Harvard Medical School, Charlestown, Massachusetts 02129, USA

⁵Department of Biosystems Science, School of Advanced Sciences, The Graduate University for Advanced Studies (Sokendai), Hayama 240-0193, Japan

⁶PRESTO, Japan Science and Technology Agency (JST), Kawaguchi 332-0012, Japan

* Present address: Novartis Institute for Biomedical Research, Cambridge, Massachusetts 02139, USA

Imprinted genes are epigenetically marked during gametogenesis so that they are exclusively expressed from either the paternal or the maternal allele in offspring¹. Imprinting prevents parthenogenesis in mammals and is often disrupted in congenital malformation syndromes, tumours and cloned animals¹. Although *de novo* DNA methyltransferases of the Dnmt3 family are implicated in maternal imprinting², the lethality of *Dnmt3a* and *Dnmt3b* knockout mice³ has precluded further studies. We here report the disruption of *Dnmt3a* and *Dnmt3b* in germ cells, with their preservation in somatic cells, by conditional knockout technology⁴. Offspring from *Dnmt3a* conditional mutant females die *in utero* and lack methylation and allele-specific expression at all maternally imprinted loci examined. *Dnmt3a* conditional mutant males show impaired spermatogenesis and lack methylation at two of three paternally imprinted loci examined in spermatogonia. By contrast, *Dnmt3b* conditional mutants and their offspring show no apparent phenotype. The phenotype of

Dnmt3a conditional mutants is indistinguishable from that of *Dnmt3L* knockout mice^{2,5}, except for the discrepancy in methylation at one locus. These results indicate that both *Dnmt3a* and *Dnmt3L* are required for methylation of most imprinted loci in germ cells, but also suggest the involvement of other factors.

We used the Cre/loxP system⁴ to generate conditional mutant mice. A conditional allele of *Dnmt3a* (referred to as *Dnmt3a*^{2lox}) was generated by gene targeting in embryonic stem (ES) cells (Fig. 1a). Cre-mediated deletion of the exons flanked by loxP sites would lead to out-of-frame splicing of the messenger RNAs coding for all active isoforms (designated as *Dnmt3a*^{1lox}). After chimera formation and germline transmission, the mice were crossed with tissue non-specific alkaline phosphatase (TNAP)-Cre knock-in mice, which express the Cre recombinase in primordial germ cells from embryonic day 9.5 (E9.5) to late gestation⁶.

The [*Dnmt3a*^{2lox/1lox}, TNAP-Cre] conditional mutant mice were smaller than their littermates by 20% in early postnatal stages (3.14 ± 0.59 g (±s.d., *n* = 21) compared with 3.78 ± 0.69 g (*n* = 52) at postnatal day 7 (P7)), probably because of Cre-mediated deletion in somatic cells: all somatic tissues examined contained a proportion of *Dnmt3a*^{1lox/1lox} cells (from 56% to 80%). Consistent with this was our observation of a leaky Cre expression in somatic tissues of early postimplantation embryos (data not shown). However, the [*Dnmt3a*^{2lox/1lox}, TNAP-Cre] mice were viable, reached normal body weight by about 4 weeks of age and survived to adulthood.

It was previously shown that paternal methylation imprints are established during the gonocyte (or prospermatogonia) stage (E14.5 to newborn)^{7,8} and maternal methylation imprints (and functional imprints) are established during the oocyte growth stage (P5–20) (refs 9, 10). We found that most E14.5 germ cells have already undergone Cre-mediated deletion in both males and females (Fig. 1b). The observed *Dnmt3a*^{2lox} signal might have arisen from contaminating somatic cells, because the average purity of our fetal germ cell preparations was 90%. In oogenesis, the *Dnmt3a*^{2lox} allele became undetectable before P3 (Fig. 1b). We were unable to examine postnatal male germ cells owing to impaired spermatogenesis (see below).

When the [*Dnmt3a*^{2lox/1lox}, TNAP-Cre] females were crossed with wild-type males, no live pups were obtained. Subsequent studies on the plugged females revealed that the embryos are slightly

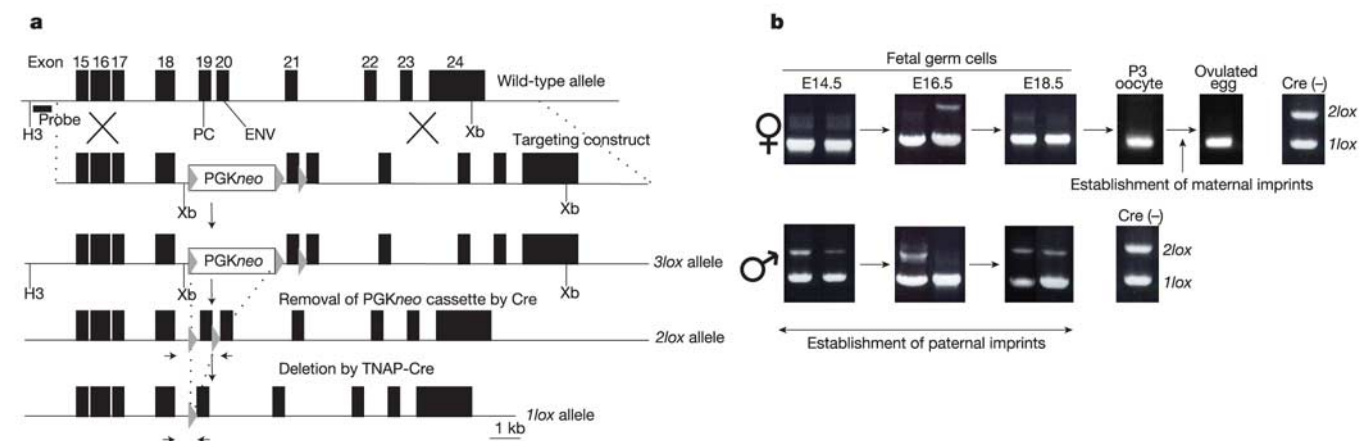


Figure 1 Conditional knockout of the *Dnmt3a* locus in the germline by TNAP-Cre. **a**, Schematic presentation of the knockout strategy. Exons are shown as filled boxes and loxP sites as hatched triangles. Positions of the highly conserved PC and ENV motifs of the methyltransferase domain are also indicated. The probe used to identify correctly targeted ES cells³ is shown as a horizontal bar and the PCR primers used to genotype mice and cells as arrowheads. Restriction sites used to diagnose correct recombinations were *Hin*

dIII (*H3*) and *Xba*I (*Xb*). **b**, Efficient deletion of *Dnmt3a* by TNAP-Cre in developing germ cells. Germ cells were isolated from [*Dnmt3a*^{2lox/1lox}, TNAP-Cre] mutant males and females and genotyped with the above primers, which can amplify the 2lox and 1lox alleles simultaneously. The timings of the establishment of paternal and maternal methylation imprints in wild-type mice are also shown.

smaller than normal embryos at E9.5 (seven litters, $n = 56$) and already dead, or almost so, at E10.5 (three litters, $n = 21$). Only dead embryos or resorptions were found at E11.5 (three litters, $n = 25$). Embryos recovered at E10.5 showed developmental defects such as open neural tube, a lack of branchial arches, and pale skin

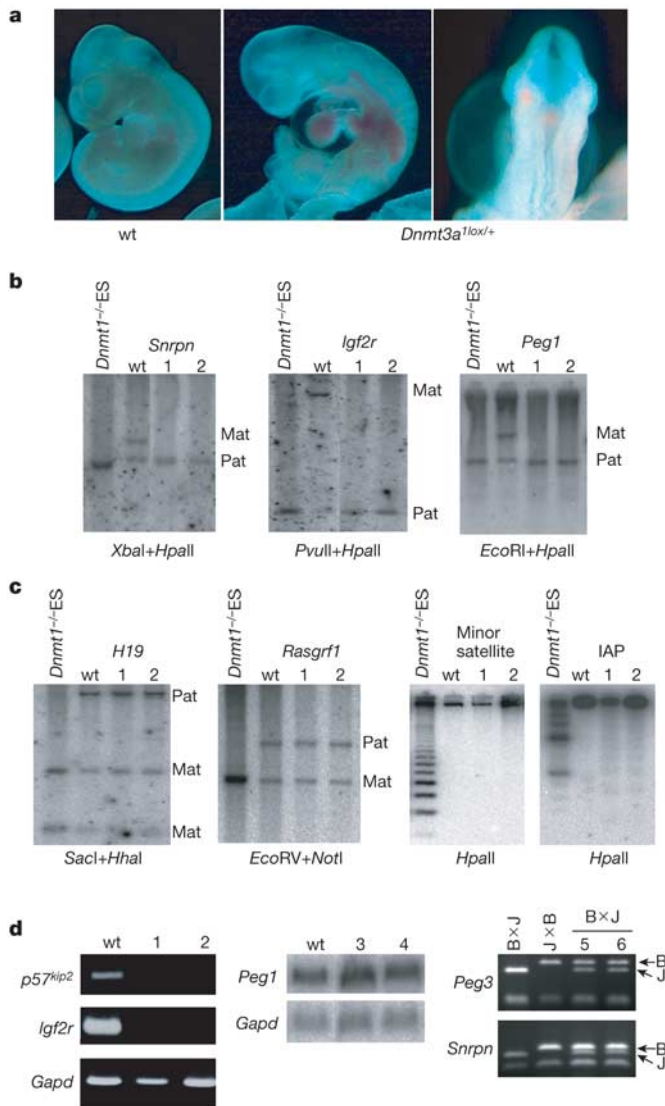


Figure 2 Developmental defects and loss of maternal imprints in embryos derived from *Dnmt3a* conditional mutant mothers. **a**, Morphology of the E10.5 embryos derived from the [*Dnmt3a*^{2lox/1lox}, TNAP-Cre] mutant females (middle and right) and a control wild-type embryo (wt, left). The embryos show defects such as smaller branchial arches, a smaller brain and an open neural tube. **b**, Disruption of methylation at three maternally imprinted loci in E10.5 embryos shown by Southern blot hybridization. The maternally derived DMR from each locus is methylated in a wild-type embryo but unmethylated in two affected embryos (1 and 2). **c**, No change in methylation at two paternally imprinted loci, minor satellite repeats and IAP repeats. Methylation was analysed by Southern blot hybridization. **d**, Disruption of expression at five maternally imprinted loci in E10.5 embryos. Expression of *p57^{kip2}* and *Igf2r* is extinct in affected embryos (1 and 2) as shown by RT-PCR, whereas expression of *Peg1* is increased in affected embryos (3 and 4) as shown by Northern blot hybridization. *Gapd* is a non-imprinted housekeeping control (encoding glyceraldehyde3phosphate dehydrogenase). *Peg3* and *Snrpn* were expressed biallelically in two affected embryos (5 and 6), as revealed by allele-specific RT-PCR. Expressed alleles were distinguished by strain-specific polymorphisms in F₁ embryos produced by crossing [*Dnmt3a*^{2lox/1lox}, TNAP-Cre] females with JF1 males. Arrows indicate the positions of bands representing the respective alleles (B, C57BL/6; J, JF1).

probably due to impediment of circulation (Fig. 2a). All recovered embryos were of *Dnmt3a*^{1lox/+} genotype, indicating a high rate of deletion by Cre. Note that this is a maternal-effect mutation, because these embryos inherited a functional *Dnmt3a* allele from their father. We also note that the defects are indistinguishable from those of the previously reported maternal-effect mutation of *Dnmt3L*^{2,5}.

Imprinted genes have differentially methylated regions (DMRs) in the 5'-flanking region or in introns¹¹. We examined the methylation status of the DMRs and other sequences in the E10.5 embryos derived from the mutant mothers. The DMRs normally methylated on the maternal allele, such as those of *Snrpn*, *Igf2r* and *Peg1*, were unmethylated (Fig. 2b). By contrast, the methylation status of the paternally methylated *H19* and *Rasgrf1* DMRs, minor satellite DNA and intracisternal A particle (IAP) repeats (endogenous retroviruses) was unaffected (Fig. 2c). However, bisulphite methylation sequencing¹² revealed that IAP sequences are about 50% less methylated in the mutant oocytes (see Supplementary Information). The results indicate that not only maternally methylated DMRs but also other sequences are affected during oogenesis and that methylation of IAP, but not the imprinted genes, can be restored after fertilization.

We then examined the expression of maternally methylated imprinted genes in the same embryos and found that *p57^{kip2}* and *Igf2r* are silenced (Fig. 2d), which is consistent with a loss of expression from the normally active maternal alleles. In addition, expression of *Peg1* was increased (1.6-fold) (Fig. 2d), indicating a derepression of the normally silent maternal allele. Allele-specific expression analyses with strain-specific polymorphisms showed that *Peg3* and *Snrpn* are expressed biallelically. Thus, functional imprints are lost at the maternally methylated loci.

When the [*Dnmt3a*^{2lox/1lox}, TNAP-Cre] males were crossed with wild-type females, no sign of pregnancy was seen in the plugged females. Examination of the mutant males revealed that their testis weight is greatly reduced in comparison with the controls (19.7 ± 1.6 mg ($\pm$ s.d., $n = 7$) versus 96.5 ± 7.9 mg ($n = 5$) at 11 weeks of age; Fig. 3a). Histological examinations showed that the testes from the mutant males at P11 contain a slightly reduced number of spermatogonia in the seminiferous tubules but that those at 11 weeks of age contain only few spermatogonia and no spermatocytes, spermatids or spermatozoa, which are abundantly present in the wild-type testes (Fig. 3a). This indicates that *Dnmt3a* is essential for spermatogenesis. The defects are again indistinguishable from those of the *Dnmt3L*^{-/-} males^{2,5}.

Because of the azoospermia, we could not ask whether the spermatozoa or offspring carry normal paternal imprints. Similarly, in the previously reported *Dnmt3L*^{-/-} males, effects of the mutations on the establishment of paternal imprints have not been addressed^{2,5}. We took advantage of laser-microdissection technology to collect spermatogonia from histological sections of P11 testes from the *Dnmt3a* conditional mutant males and *Dnmt3L*^{-/-} males. We analysed three paternally methylated DMRs by bisulphite sequencing and found that the normally methylated *H19* DMR is unmethylated in both mutants (Fig. 3b). Interestingly, normally methylated IG-DMR at *Dlk1-Gtl2* was methylated in *Dnmt3L*^{-/-} spermatogonia but not in [*Dnmt3a*^{2lox/1lox}, TNAP-Cre] spermatogonia (Fig. 3b). Both *Rasgrf1* DMR and IAP sequences were methylated in these mutants as well as in wild-type spermatogonia (Fig. 3b, and see Supplementary Information). Thus, *Dnmt3a* is essential for the methylation of two of the three paternally imprinted loci examined.

A conditional *Dnmt3b* allele (referred to as *Dnmt3b*^{2lox}) was also produced, of which details are described elsewhere (J. Dodge, M.O., Y. Ueda and E.L., in preparation). We generated [*Dnmt3b*^{2lox/1lox}, TNAP-Cre] mice, which were phenotypically normal. Analyses of germ cells at various stages revealed that a significant proportion of the E14.5 germ cells have already undergone deletion in both male

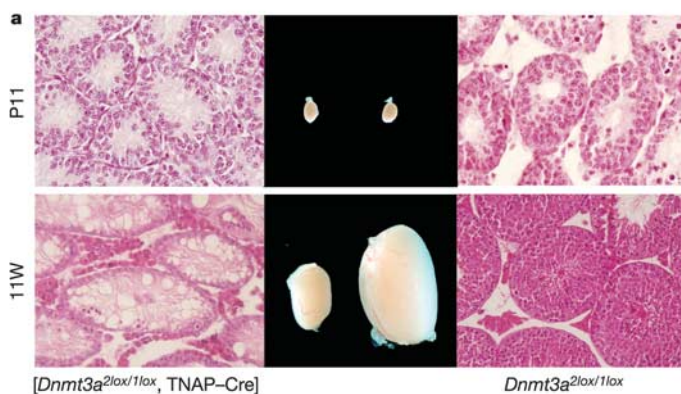
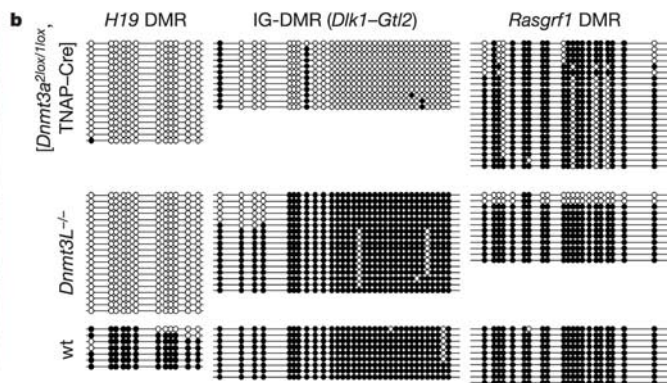


Figure 3 Impaired spermatogenesis and failure of establishment of paternal methylation imprints in *Dnmt3a* conditional mutant males. **a**, Gross morphology and histology of testes from *[Dnmt3a^{2lox/1lox}, TNAP-Cre]* mutant males and control *Dnmt3a^{2lox/1lox}* males at P11 and 11 weeks (11W) after birth. The number of spermatogonia at P11 is not very different between the affected and control testes. At 11 weeks of age, the testis size is greatly reduced in the mutant males and no spermatocytes, spermatids or spermatozoa are seen

and females (Fig. 4). In oogenesis, the *Dnmt3b^{2lox}* allele became undetectable before P3 (Fig. 4). In spermatogenesis, Cre-mediated deletion was nearly completed by the newborn stage (Fig. 4).

When the *[Dnmt3b^{2lox/1lox}, TNAP-Cre]* males and females were crossed with wild-type partners, healthy pups were obtained. All 102 pups obtained from three male mutants (a total of 15 litters; average litter size 6.8) were of *Dnmt3b^{1lox/+}* genotype, indicating a 100% deletion rate. Among the 88 pups obtained from four female mutants (13 litters; average litter size 6.8), 87 were of the *Dnmt3b^{1lox/+}* genotype, indicating a 98% deletion rate. We analysed the methylation status of imprinted genes and other sequences, but all was found to be normal (data not shown). Notably, *Rasgrf1*, which was unaffected in *[Dnmt3a^{2lox/1lox}, TNAP-Cre]* spermatogonia, was also normally methylated in *[Dnmt3b^{2lox/1lox}, TNAP-Cre]* spermatozoa, implying functional redundancy of *Dnmt3a* and *Dnmt3b* on this locus.

Last, to confirm that our *Dnmt3a^{1lox}* and *Dnmt3b^{1lox}* alleles were functionally null, homozygous mutants were generated. The *Dnmt3a^{1lox/1lox}* and *Dnmt3b^{1lox/1lox}* mutants showed phenotypes indistinguishable from those of previously described *Dnmt3a^{-/-}* and *Dnmt3b^{-/-}* mice, respectively³. In fact, no *Dnmt3a* and *Dnmt3b* proteins were detected in the *Dnmt3a^{1lox/1lox}* and *Dnmt3b^{1lox/1lox}* mutants (data not shown), respectively, perhaps



in their seminiferous tubules. **b**, Methylation analysis at three paternally imprinted loci in spermatogonia from *Dnmt3a* conditional mutant males and *Dnmt3L^{-/-}* males. Spermatogonia were collected from P11 testes of *[Dnmt3a^{2lox/1lox}, TNAP-Cre]*, *Dnmt3L^{-/-}* (ref. 2), and wild-type (wt) males by laser microdissection. Methylation of the three DMRs was studied by bisulphite genomic sequencing. Circles represent individual CpG dinucleotides. Filled circles, methylated; open circles, unmethylated.

because of quick degradation of the aberrant transcripts or proteins.

Our findings clearly indicate that *Dnmt3a* is required for the establishment of both maternal and paternal imprints. The role of *Dnmt3a* in maternal imprinting is consistent with its expression in growing oocytes¹⁰ and with our previous observation that embryos derived from a *[Dnmt3a^{-/-}, Dnmt3b^{+/-}]* ovary transplanted into a normal female lacked the maternal imprints². Furthermore, this study provides the first evidence that *Dnmt3a* is required for the establishment of paternal methylation imprints. This is consistent with our recent finding that *Dnmt3a* is expressed in gonocytes (ref. 13 and N.T., T. Chen, E.L. and H.S., unpublished observations).

The defects observed in the *Dnmt3a* conditional mutants were surprisingly similar to those of the *Dnmt3L* mutants^{2,5}. We have previously shown that these two proteins interact in transfected cells². Furthermore, human DNMT3L has been shown to enhance the *de novo* methylation activity of *Dnmt3a* (ref. 14), which suggests that *Dnmt3L* is an important cofactor for *Dnmt3a*. Although we observed a discrepancy at one paternally methylated locus, the identification of *Dnmt3a* and *Dnmt3L* as common components of the *de novo* methylation activity in both germlines is consistent with the acquisition of a paternal methylation imprint in the female germline by an imprinted locus involving chromosomal inversion¹⁵ and indicates that one or more additional factors are required for gamete-specific differential methylation. A recessive human disorder, familial biparental complete hydatidiform mole, which specifically and globally affects the maternal imprints but has no mutations at DNMT1o, DNMT3A, DNMT3B or DNMT3L (refs 16–19), indicates the presence of such factors.

In conclusion, our findings strongly suggest that *de novo* methylation is involved in the initiation of imprinting in mammals. This contrasts with our recent finding that *de novo* methylation is dispensable for the initiation of monoallelic expression (or silencing) of the *Xist* locus during random X-chromosome inactivation²⁰. Despite the presence of several similarities between the two phenomena, they seem to use different epigenetic mechanisms for their initiation. □

Methods

Gene targeting and mice

A targeting vector for *Dnmt3a* was electroporated into J1 ES cells, which were subsequently selected for integration in G418-containing medium as described²¹. Correctly targeted ES cells (with a *3lox* allele) were identified by Southern-blot analysis³ (Fig. 1a). The floxed region contained exon 19, which codes for the conserved PC motif of the catalytic domain. The ES cells were injected into blastocysts to generate germline chimaeras. The selectable

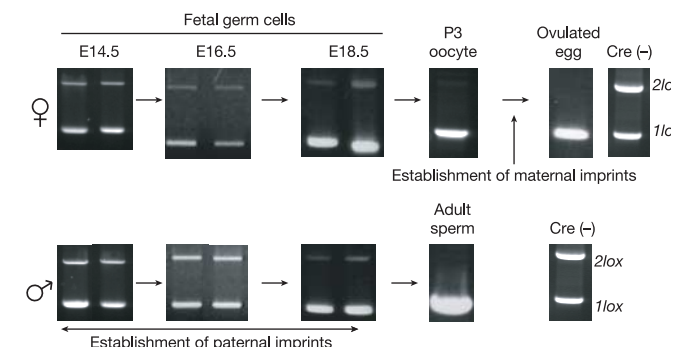


Figure 4 Efficient deletion of *Dnmt3b* by TNAP-Cre in developing germ cells. Germ cells were isolated from *[Dnmt3b^{2lox/1lox}, TNAP-Cre]* mutant males and females and genotyped using the primers that can amplify the *2lox* and *1lox* alleles simultaneously. The timings of the establishment of paternal and maternal methylation imprints in wild-type mice are also shown.

drug-resistance marker was removed in mice by crossing with EIIa-Cre mice²² to produce the *2lox* allele (referred to as *Dnmt3a*^{2lox}). Production of the conditional *Dnmt3b* allele (referred to as *Dnmt3b*^{2lox}) is described elsewhere (J. Dodge, M.O., Y. Ueda and E.L., in preparation). To inactivate *Dnmt3a* and *Dnmt3b* preferentially in primordial germ cells, we crossed these mice with TNAP-Cre knock-in mice⁶. Mice were genotyped by polymerase chain reaction (PCR; primer sequences are available from the authors on request).

Preparation of germ cells

Fetal germ cells were collected as described²³. Oocytes from females at P3 were obtained in accordance with the published protocol²⁴. Ovulated oocytes and epididymal sperm were isolated by a standard procedure. Spermatogonia were isolated from histological sections of P11 testes by laser microdissection with the Robot MicroBeam System (PALM).

Isolation of DNA and RNA from tissues and cells

DNA was prepared from tail biopsies, whole embryos, various adult organs and germ cells including spermatogonia collected by laser microdissection in accordance with the standard protocol. Total RNA was isolated from whole embryos with ISOGEN (Nippon Gene) in accordance with the protocol provided by the manufacturer.

DNA methylation analysis

Genomic DNA (5 µg) was digested with methylation-sensitive and/or methylation-insensitive enzymes, and analysed by Southern blot hybridization. The probes used were as follows: a 2.5-kilobase (kb) fragment for IAP repeats, a 0.1-kb insert of pMR150 for centromeric minor satellite DNA, a 2.0-kb fragment from the *H19* DMR, a 0.7-kb Sp-4 repeat from the *Rasgrf1* DMR, a 1.1-kb fragment from the *Igf2r* DMR2, a 1.4-kb fragment from intron 1 of *Peg1*, and a 2.2-kb fragment from the *Snrpn* DMR1. Bisulphite sequencing analysis was performed with the EZ DNA Methylation Kit (Zymo Research). The primer sequences and PCR conditions are available from the authors on request.

Expression analysis

Complementary DNA was synthesized from 1 µg of total RNA with Superscript II reverse transcriptase (Life Technologies) with random primers. PCR was performed with gene-specific primers: primer sequences and PCR conditions are available from the authors on request. For the allele-specific expression analysis, restriction-fragment-length polymorphisms were identified between the laboratory strains and JF1. Amplified cDNA fragments of *Peg3* and *Snrpn* were digested with *TaqI* and *NlaIII*, respectively. For northern blot analysis, total RNA (10 µg) was fractionated on an agarose-formaldehyde gel, and transferred to a piece of Biodyne B nylon membrane (Pall) in accordance with standard procedures. RNA was crosslinked by ultraviolet and hybridized to ³²P-labelled probes. Probes used were a 1.0-kb *Peg1* cDNA and a 0.4-kb *Gapd* cDNA. The latter cDNA was amplified by PCR with reverse transcription (RT-PCR). Primer sequences are available from the authors on request.

Histology

Tissues were fixed with Bouin's solution (Muto Pure Chemicals), dehydrated and embedded in paraffin. Sections 6 µm thick were prepared with a microtome and stained with haematoxylin and eosin.

Received 19 March; accepted 21 April 2004; doi:10.1038/nature02633.

1. Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nature Rev. Genet.* **2**, 21–32 (2001).
2. Hata, K., Okano, M., Lei, H. & Li, E. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development* **129**, 1983–1993 (2002).
3. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247–257 (1999).
4. Torres, R. M. & Kühn, R. *Laboratory Protocols for Conditional Gene Targeting* (Oxford Univ. Press, 1997).
5. Bourc'his, D., Xu, G. L., Lin, C. S., Bollman, B. & Bestor, T. H. Dnmt3L and the establishment of

maternal genomic imprints. *Science* **294**, 2536–2539 (2001).

6. Lomeli, H., Ramos-Mejia, V., Gertsenstein, M., Lobe, C. G. & Nagy, A. Targeted insertion of Cre recombinase into the TNAP gene: excision in primordial germ cells. *Genesis* **26**, 116–117 (2000).
7. Ueda, T. *et al.* The paternal methylation imprint of the mouse H19 locus is acquired in the gonocyte stage during foetal testis development. *Genes Cells* **5**, 649–659 (2000).
8. Davis, T. L., Yang, G. J., McCarrey, J. R. & Bartolomei, M. S. The H19 methylation imprint is erased and re-established differentially on the parental alleles during male germ cell development. *Hum. Mol. Genet.* **9**, 2885–2894 (2000).
9. Obata, Y. & Kono, T. Maternal primary imprinting is established at a specific time for each gene throughout oocyte growth. *J. Biol. Chem.* **277**, 5285–5289 (2002).
10. Lucifero, D., Mann, M. R. W., Bartolomei, M. S. & Trasler, J. M. Gene-specific timing and epigenetic memory in oocyte imprinting. *Hum. Mol. Genet.* **13**, 839–849 (2004).
11. Ferguson-Smith, A. C. & Surani, M. A. Imprinting and the epigenetic asymmetry between parental genomes. *Science* **293**, 1086–1089 (2001).
12. Clark, S. J., Harrison, J., Paul, C. L. & Frommer, M. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.* **22**, 2990–2997 (1994).
13. Chen, T., Ueda, Y., Xie, S. & Li, E. A novel Dnmt3a isoform produced from an alternative promoter localizes to euchromatin and its expression correlates with active *de novo* methylation. *J. Biol. Chem.* **277**, 38746–38754 (2002).
14. Chedin, F., Lieber, M. R. & Hsieh, C. L. The DNA methyltransferase-like protein DNMT3L stimulates *de novo* methylation by Dnmt3a. *Proc. Natl Acad. Sci. USA* **99**, 16916–16921 (2002).
15. Cerrato, F. *et al.* Paternal imprints can be established on the maternal Igf2–H19 locus without altering replication timing of DNA. *Hum. Mol. Genet.* **12**, 3123–3132 (2003).
16. Fisher, R. A. *et al.* The maternally transcribed gene p57^{KIP2} (CDNK1C) is abnormally expressed in both androgenetic and biparental complete hydatidiform moles. *Hum. Mol. Genet.* **11**, 3267–3272 (2002).
17. Judson, H., Hayward, B. E., Sheridan, E. & Bonthron, D. T. A global disorder of imprinting in the human female germ line. *Nature* **416**, 539–542 (2002).
18. El-Maari, O. *et al.* Maternal alleles acquiring paternal methylation patterns in biparental complete hydatidiform moles. *Hum. Mol. Genet.* **12**, 1405–1413 (2003).
19. Hayward, B. E. *et al.* Lack of involvement of known DNA methyltransferases in familial hydatidiform mole implies the involvement of other factors in establishment of imprinting in the human female germline. *BMC Genet.* **4**, 2 (2003).
20. Sado, T., Okano, M., Li, E. & Sasaki, H. De novo DNA methylation is dispensable for the initiation and propagation of X chromosome inactivation. *Development* **131**, 975–982 (2004).
21. Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915–926 (1992).
22. Xu, X. *et al.* Direct removal in the mouse of a floxed neo gene from a three-loxP conditional knockout allele by two novel approaches. *Genesis* **30**, 1–6 (2001).
23. Hashimoto, N., Kubokawa, R., Yamazaki, K., Noguchi, M. & Kato, Y. Germ cell deficiency causes testis cord differentiation in reconstituted mouse fetal ovaries. *J. Exp. Zool.* **253**, 61–70 (1990).
24. Bao, S., Obata, Y., Carroll, J., Domeki, I. & Kono, T. Epigenetic modifications necessary for normal development are established during oocyte growth in mice. *Biol. Reprod.* **62**, 616–621 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Nagy and H. Lomeli (Samuel Lunenfeld Research Institute) for providing the TNAP-Cre mice; K. Shiota and S. Tanaka (The University of Tokyo) for help in laser-microdissection microscopy; K. Kumaki and Y. Kato for help and advice on bisulphite sequencing; and C. Suda, M. Kanbayashi, M. Serizawa and H. Furuumi for technical assistance and mouse maintenance. This work was supported in part by Grants-in-Aid from the Ministry of Education, Culture, Sports, Science, and Technology of Japan to H.S., and grants from the National Institutes of Health to E.L.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.S. (hisasaki@lab.nig.ac.jp). Requests for the Dnmt3a conditional knockout mice should be addressed to E.L. (en.li@pharma.novartis.com).

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)
Andy Douglas (4975)
Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequinot (4974)

Production Manager:

Billie Franklin
To send materials use London
address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffen
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mail: p.phelan@nature.com
o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: rasami@naturejp.com

naturejobs

Stem-cell state lines

Research involving human embryonic stem cells has sparked heated debate around the world, and is currently at the heart of a complex balancing act in the United States. On the one hand it is responsible for one of the largest US recruitment drives in recent years. But this surge in interest is tempered by politics and demarcated by state lines.

US law prohibits federal funds from being used on research that involves newly derived embryonic stem-cell lines. Private funds, however, do not suffer from such restrictions, and individual states are now trying to decide whether to plough their money into such an enterprise. Recent events in Missouri and Kansas, which both want to increase their national research presence, are cases in point.

In Missouri last month, university officials and biotech leaders had hoped that a state economic-development package would allow universities to combine private funds with public money to create endowed chairs. But this provision was quashed, in part because anti-abortion groups voiced concern that the money could be used to support embryonic stem-cell research.

Kansas faced a similar situation. Here, in April, the state passed a law to create a bioscience authority that would allow it to identify and recruit top researchers — but that basically bars embryonic stem-cell research.

The next battleground will be California, which already leads the country in terms of receiving funding from the National Institutes of Health. The California Stem Cell Research and Cures Initiative will provide \$3 billion in stem-cell funding over ten years. So it will be up to California's voters in this November's referendum to tip the recruiting balance further towards their state by agreeing to more stem-cell funding, without restrictions. In the end, it may be up to researchers to follow the funding.

Paul Smaglik
Naturejobs editor



Contents

CAREER VIEW

Bricks & Mortar

Scandinavia's material gains

Graduate Journal

Show us the money!

Movers

Axel Ullrich

p906

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

GRADUATE JOURNAL

Show us the money!

After attending the Biotechnology Industry Organization conference in San Francisco this month, I realized that a lot of modern biomedical research hinges on money. The more cash you have, the more science you can do.

Although the pace of scientific discovery can be slow, the creativity that drives it is not, with the best ideas for new research directions often coming quite early in one's career. I would bet that many Nobel laureates made their breakthroughs before the age of 42 — the average age today's PhDs get their first academic appointment.

This leaves young researchers with the challenge of getting maximum resources at the earliest possible stage of their career. But in academic science, seniority tends to rule the roost. Occasionally, visionary mentors will recognize and fully support a talented young investigator proposing a new approach to a long-standing scientific question. But these scenarios are not the norm, and many good scientific ideas must have been lost or delayed because their originator's career had not yet blossomed.

So in addition to thinking up new ways to drive future scientific discovery, young scientists will have to be even more creative to convince the world to 'show us the money' (or give us the resources) to put our new ideas in play. ■

Tshaka Cunningham is a fifth-year graduate student at Rockefeller University in New York.

BRICKS & MORTAR

Scandinavia's material gains

With the launch this month of two virtual centres dedicated to nanotechnology, Scandinavia is staking its claim as a hot spot for materials science. Based in Denmark and Norway, the facilities look set to create research and training positions for several hundred materials scientists.

In Copenhagen, the Technical University of Denmark unveiled NANO•DTU, which will be one of Europe's largest nanoscience centres. Currently it employs about 100 researchers, including postdocs and 40 PhD students. But this is likely to increase to help the university compete for the anticipated rise in government funds for nanotechnology, says physicist Jens Nørskov, the centre's director. "We are

gearing up so that we will present ourselves as a good place to invest, should the Danish government choose to invest more in nanotechnology," he says.

The chief physical attraction at the centre is the newly built DANCHIP, a 1,000-square-metre clean room. Equipped with atomic force microscopy and nanolithography tools, the room will allow researchers to etch nanomachines as small as 10 nanometres.

Nørskov says that the new facilities, combined with a new BSc in nanotechnology, should help to attract undergraduates who lately have been shying away from physics and chemistry. "We're trying to make our curriculum interesting to young people," he says.

And in Norway, the University of Oslo has opened the doors on its Centre for Materials Science and

Nanotechnology (SMN). Arising as the result of a reorganization in the faculty of mathematics and natural sciences, the centre features six research groups across four buildings.

The most tangible investment in the virtual centre is the Micro- and Nanotechnology laboratory (MINA-lab), a Nkr232-million (US\$33.7-million) 6,000-square-metre facility, with a clean room.

The SMN will employ 18 permanent staff, but with technicians, administrators, students and postdocs, the total effort will feature some 90 people. Like the Danish initiative, the SMN aims to improve international recruitment into Scandinavian nanotech. It will also focus on using nanotechnology and functional materials for sustainable energy production, one focus of the Norwegian government's research priorities. ■

Paul Smaglik is editor of Naturejobs.

MOVERS

Axel Ullrich, visiting scientist and director, Singapore Onco Genome Laboratory



'Right place, right time' could be the title of Axel Ullrich's biography. He moved to the epicentre of genetic engineering at the field's genesis, joined one of the most successful biotechnology companies in its early stages, and is now preparing to participate in one of the largest, centralized expansions in life-science research.

Plotting a course through such events has always come easily to Ullrich. After completing his PhD at the University of Heidelberg in Germany in the early 1970s, he watched genetic engineering begin to find its feet, and he

knew exactly what to do. "I didn't even have to think about it," Ullrich says. "The United States was the place to go." Apart from being attracted by the scientific opportunities and better funding, Ullrich thought that living across the Atlantic would be a good way to learn English.

But it was not his command of a new language that impressed scientists around the world in 1977. It was his publication of the first transfer of a mammalian gene — encoding for human insulin — into bacteria, a result of two years' work at the University of California, San Francisco. "Holding the sequence of insulin in our hand for the first time was a great moment," Ullrich recalls.

A year later, Ullrich left the university to join the nascent company Genentech, a forerunner of the biotech industry. There he developed the first genetically engineered medicine, human insulin,

made by bacteria, and also got increasingly interested in cancer research. His work helped to bring the cancer drug Herceptin to the market in 1998. But initially Genentech didn't pursue the drug, so Ullrich returned to Germany as a director at the Max Planck Institute of Biochemistry near Munich. "It was not easy for me to adopt to the more rigid German scientific community," he says. "But, in retrospect, it was the right choice."

This month Ullrich took on a new challenge. He moved as a visiting scientist to Singapore's 'Biopolis' to work for A*STAR — the Agency for Science, Technology and Research, which charts the course for the city-state's science and technology. There he will set up a group to carry out large-scale genomics dedicated to cancer-gene discovery. After his year is over, it will become apparent whether he was again in the right place at the right time. ■

CV **1988–:** Director, Max Planck Institute of Biochemistry, Martinsried, Germany
1978–88: Senior scientist, Genentech, South San Francisco, California
1975–77: Postdoctoral fellow, University of California, San Francisco